

Information Recovery in Shuffled Graphs via Graph Matching

Vince Lyzinski

Johns Hopkins University Human Language Technology Center of Excellence

May 10, 2016

Abstract

In a number of methodologies for joint inference across graphs, it is assumed that an explicit vertex correspondence is *a priori* known across the vertex sets of the graphs. While this assumption is often reasonable, in practice these correspondences may be unobserved and/or errorfully observed, and graph matching—aligning a pair of graphs to minimize their edge disagreements—is used to align the graphs before performing subsequent inference. Herein, we explore the duality between the loss of mutual information due to an errorfully observed vertex correspondence and the ability of graph matching algorithms to recover the true correspondence across graphs. We then demonstrate the practical effect that graph shuffling—and matching—can have on subsequent inference, with examples from two sample graph hypothesis testing and joint graph clustering.

1 Introduction

Graphs are an increasingly popular data modality in scientific research and statistical inference, with diverse applications in connectomics [6], social network analysis [7], and pattern recognition [22], to name a few. Many joint graph inference methodologies (see, for example, [45, 19, 6, 38]), joint graph embedding algorithms (see, for example, [20, 35, 41, 40]) and graph-valued time-series methodologies (see, for example, [24, 34, 46, 50]) operate under the implicit assumption that an explicit vertex correspondence is *a priori* known across the vertex sets of the graphs. While this assumption is natural in a host of real data settings, in many applications these correspondences may be unobserved and/or errorfully observed [48]. Connectomics offers a striking example of this continuum. Indeed, while for some simple organisms (e.g., the *C. elegans* roundworm [52]) explicit neuron labels are known across specimen, and in human DTMRI connectomes, the vertices are often regions of the brain registered to a common template (see [19]), explicit cross-subject neuron labels are often unknown for more complex organisms.

In the presence of a latent vertex correspondence that is unknown or errorfully observed across graphs, graph matching methodologies can be applied to recover the latent vertex alignment before performing subsequent inference. As a result, as joint graph inference has surged in popularity, so has graph matching; see [11] and [17] for an excellent review of the graph matching literature. Formally, given two graphs G_1 and G_2 , with respective adjacency matrices A and B , the graph

matching problem (GMP) seeks to minimize $\|AP - PB\|_F$ over permutation matrices P ; i.e., the GMP seeks a relabeling of the vertices of B that minimizes the number of induced edge disagreements between A and PBP^T ; see Section 3 for more detail. While the related graph isomorphism problem has recently been shown to be of quasipolynomial complexity [5], there are no efficient algorithms known for the more general problem of graph matching. Due to its practical utility and computational difficulty, myriad heuristics have been proposed in the literature for approximately solving the GMP; see, for example, [11] and [49] and the references contained therein.

How can we quantify the effect of the added uncertainty due to an errorfully observed vertex correspondence? Heuristically, if (G_1, G_2) is a realization from a bivariate random graph model with the property that vertices that are aligned across graphs behave similarly in their respective networks, then the uncertainty in G_2 is greatly reduced by observing G_1 and the latent alignment. Indeed, in the extreme case of G_1 and G_2 being isomorphic, observing the latent alignment function and G_1 completely determines G_2 . However, as the vertex labels are shuffled uncertainty is introduced into the bivariate model. In order to formalize this heuristic, we adopt an information theoretic perspective (see [12] for the necessary background). We develop a bivariate graph model, the ρ -correlated stochastic blockmodel (Section 2.1), in which we are able to address the following questions:

- (i) What is the information loss/increase in uncertainty due to an errorful labeling across graphs?
- (ii) How does this information loss impact subsequent inference?
- (iii) Can graph matching recover the lost information?

In the process, we uncover a duality between graph *matchability* (see Definition 6) and information loss. Under mild model assumptions, in the regime where graph matching can recover the latent vertex alignment after shuffling, relatively little information is lost via shuffling. We conjecture the inverse statement to be true as well: In the regime where graph matching cannot recover the latent vertex alignment after shuffling, a relatively nontrivial amount of information is lost in the shuffle. In addition, in the aforementioned ρ -correlated SBM model, we are able to establish a phase transition for correlated graphs transitioning from being matchable to unmatchable (see Theorems 10–12), and conjecture the same phase transition for the relative information loss as well.

While in the presence of modest correlation relatively little information is lost due to shuffling, we demonstrate in a pair of inference tasks—two sample graph hypothesis testing and joint vertex clustering—that the lost information can have a dramatic negative effect on subsequent inference. We expect the same negative effect of shuffling will occur in many other inference tasks as well. While this may seem like an indictment against joint inference in the errorful correspondence setting, we also show that graph matching can effectively recover all of the lost information; see Theorem 15. This provides a theoretical basis for what we later demonstrate in our experiments: shuffling the vertex labels causes information to be lost and decreases inference performance, and graph matching recovers the true correspondence and recovers the lost performance. Together, we believe these results provide a novel theoretical understanding for the utility of graph matching across a host of inference tasks.

The paper is laid out as follows. In Section 2, we develop a bivariate random graph model endowed with a natural latent alignment between the vertex sets of the two graphs, and we rigorously define the notion of errorful/shuffled labels in our model. In Section 3 we explore the effect that vertex shuffling has on the mutual information, and explore the dual relationship between information loss and graph matchability. Lastly, in Section 4, for both real and synthetic data, we empirically explore the negative effect of vertex shuffling on two sample hypothesis testing for graphs and joint vertex clustering, and we demonstrate the capacity of graph matching to ameliorate the inferential effect of shuffled vertices. The proofs of the key results are found in Appendix A.

Note: Throughout, for real-valued function $f(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ and $g(\cdot) : \mathbb{R} \mapsto \mathbb{R}$, we shall write $f(n) \sim g(n)$ if $\lim_{n \rightarrow \infty} f(n)/g(n) = 1$. We will also make use of the abbreviation a.a.s. (for asymptotically almost surely) which will be used as follows. A sequence of events E_n occurs a.a.s. if $\mathbb{P}(E_n^c) \rightarrow 0$ at a rate fast enough to ensure $\sum_n \mathbb{P}(E_n^c) < \infty$.

2 Background

In order to quantify the information loss due to the vertex correspondence across graphs being unknown or errorfully known, we first develop a joint graph model amenable to this analysis: correlated stochastic blockmodel (SBM) random graphs [30].

2.1 Correlated Stochastic Blockmodels

SBM random graphs are widely used to model networks exhibiting an underlying community structure [21, 51]. Although they are often an overly simplistic model for real network data sets, SBMs provide a simple class of random graph models which has been effectively used to approximate the behavior of larger, more complex data sets [2, 53, 10].

Letting \mathcal{G}_n be the set of labeled, n -vertex, simple, undirected graphs, we define

Definition 1. *Two n -vertex random graphs $(G_1, G_2) \in \mathcal{G}_n \times \mathcal{G}_n$ are ρ -correlated SBM(K, \vec{n}, b, Λ) graphs (abbreviated ρ -SBM) if:*

1. $G_1 = (V, E(G_1))$ and $G_2 = (V, E(G_2))$ are marginally SBM(K, \vec{n}, b, Λ) random graphs; i.e., for each $i = 1, 2$,

- i. *The vertex set V is the union of K blocks V_1, V_2, \dots, V_K , which are disjoint sets with respective cardinalities n_1, n_2, \dots, n_K ;*
- ii. *The block membership function $b : V \mapsto [K] = \{1, 2, \dots, K\}$ is such that for each $v \in V$, $b(v)$ denotes the block of v ; i.e., $v \in V_{b(v)}$;*
- iii. *The block adjacency probabilities are given by the symmetric matrix $\Lambda \in [0, 1]^{K \times K}$; i.e., for each pair of vertices $\{j, \ell\} \in \binom{V}{2}$, the adjacency of j and ℓ is an independent Bernoulli trial with probability of success $\Lambda_{b(j), b(\ell)}$.*

2. *The random variables*

$$\{\mathbb{1}[\{j, k\} \in E(G_i)]\}_{i=1,2; \{j,k\} \in \binom{V}{2}}$$

are collectively independent except that for each $\{j, k\} \in \binom{V}{2}$, the correlation between $\mathbb{1}[\{j, k\} \in E(G_1)]$ and $\mathbb{1}[\{j, k\} \in E(G_2)]$ is $\rho \geq 0$.

Given $(G_1, G_2) \sim \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$ with $\rho \in (0, 1)$ (so that G_1 and G_2 are a.a.s. *not* isomorphic), this construction highlights the natural alignment between the vertex sets of the graphs: namely the identity function $\text{id}_n : [n] \mapsto [n]$. Indeed, for modest ρ (see Theorem 12), with high probability the identity function is the permutation of the vertex set of G_2 that best preserves the shared structure between G_1 and G_2 ; i.e., if the respective adjacency matrices of G_1 and G_2 are A and B , then with high probability $\arg\min_{P \in \Pi(n)} \|AP - PB\|_F = \{I_n\}$ where $\Pi(n)$ is the set of $n \times n$ permutation matrices. As, in practice, this alignment is often unobserved, we shall refer to id_n as the *latent alignment* between G_1 and G_2 .

One of the keys to the theoretical tractability of the ρ -SBM model is that we can construct $\rho\text{-SBM}(K, \vec{n}, b, \Lambda)$ random graphs (G_1, G_2) as follows:

1. Generate G_1 from the underlying $\text{SBM}(K, \vec{n}, b, \Lambda)$ model;
2. For each $\{j, \ell\} \in \binom{V}{2}$, $\mathbb{1}[\{\ell, j\} \in E(G_2)]$ is an independent Bernoulli trial with probability of success $\Lambda_{b(j), b(\ell)} + \rho(1 - \Lambda_{b(j), b(\ell)})$ if j and ℓ are adjacent in G_1 , and probability of success $\Lambda_{b(j), b(\ell)}(1 - \rho)$ if j and ℓ are not adjacent in G_1 .

We will leverage this construction throughout the proofs of our main results; see Appendix A for detail.

Remark 2. Note that ρ -correlated Erdős-Rényi(n, p) ($\rho\text{-ER}(n, p)$) random graphs are easily realized by letting $K = 1$ in Definition 1.

2.2 Shuffled ρ -correlated SBM random graphs

Given graph-valued random variables, $(G_1, G_2) : \Omega \mapsto \mathcal{G}_n \times \mathcal{G}_n$, the mutual information of G_1 and G_2 is defined in the standard way via

$$I(G_1; G_2) = \sum_{x, y \in \mathcal{G}_n} \mathbb{P}(G_1 = x, G_2 = y) \log \left(\frac{\mathbb{P}(G_1 = x, G_2 = y)}{\mathbb{P}(G_1 = x)\mathbb{P}(G_2 = y)} \right),$$

and the entropy of G_1 is defined in the standard way via

$$H(G_1) = - \sum_{x \in \mathcal{G}_n} \mathbb{P}(G_1 = x) \log(\mathbb{P}(G_1 = x)). \quad (1)$$

If $\rho = 0$, then two ρ -correlated SBM random graphs are independent, and the mutual information between them is 0, regardless of whether the latent vertex alignment is known across graphs or not. If $\rho = 1$, then G_1 and G_2 are isomorphic and $I(G_1; G_2) = H(G_1) = H(G_2)$, the entropy of G_1 . If $\rho > 0$, then there is nontrivial information shared across graphs, information which is potentially lost if the labeling is unknown or corrupted. Indeed, we have the following.

Proposition 3. *Let $(G_1, G_2) \sim \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$.*

i. If Λ , K and ρ are fixed in n , then

$$I(G_1; G_2) = \Theta(n^2).$$

ii. For fixed Λ and K , if $\rho \rightarrow 0$ as $n \rightarrow \infty$ then

$$I(G_1; G_2) \sim \frac{\rho^2 \binom{n}{2}}{2}.$$

We will prove Proposition 3 in Section A.1.

To understand the effect that an unobserved or errorfully observed latent alignment function has on $I(G_1; G_2)$, we first need to define the action of errorfully aligning two ρ -SBM random graphs. To this end, let S_n be the set of permutations of $[n]$. For $x = (V, E_x) \in \mathcal{G}_n$, and $\phi \in S_n$, we define the ϕ -shuffled graph $\phi(x) = (V, E_{\phi(x)}) \in \mathcal{G}_n$ via

$$\{i, j\} \in E_x \text{ iff } \{\phi(i), \phi(j)\} \in E_{\phi(x)}. \quad (2)$$

Equivalently, if the adjacency matrix of x is A_x and the permutation matrix associated with ϕ is $P_\phi \in \Pi(n)$, then the adjacency matrix of $\phi(x)$ is $P_\phi A_x P_\phi^T$.

In the context of $(G_1, G_2) \sim \rho$ -SBM, the action of a deterministic shuffling can be realized via: for all $x, y \in \mathcal{G}_n$,

$$\mathbb{P}(G_1 = x, \phi(G_2) = y) := \mathbb{P}(G_1 = x, G_2 = \phi^{-1}(y)).$$

It is not difficult to see that no information is lost from deterministically shuffling the vertices of G_2 , i.e., $I(G_1; G_2) = I(G_1; \phi(G_2))$ for $\phi \in S_n$, although subsequent inference may still be negatively impacted if ϕ is unknown. We next define the action of a random shuffling on the vertices of ρ -SBM random graphs.

Definition 4. Let σ be an S_n -valued random variable. The random graphs $(G_1, \sigma(G_2)) \in \mathcal{G}_n \times \mathcal{G}_n$ are σ -shuffled, ρ -correlated SBM(K, \vec{n}, b, Λ) graphs (abbreviated σ, ρ -SBM) if

i. (G_1, G_2) are ρ -SBM(K, \vec{n}, b, Λ) random graphs

ii. For any $x, y \in \mathcal{G}_n \times \mathcal{G}_n$, we have

$$\mathbb{P}(G_1 = x, \sigma(G_2) = y) = \sum_{\phi \in S_n} \mathbb{P}(\sigma = \phi^{-1}) \mathbb{P}(G_1 = x, G_2 = \phi(y)).$$

Definition 4 defines the observed data likelihood of $\mathbb{P}(G_1 = x, \sigma(G_2) = y)$ by summing over the total data likelihood given by

$$\mathbb{P}(G_1 = x, \sigma(G_2) = y, \sigma = \phi) = \mathbb{P}(\sigma = \phi) \mathbb{P}(G_1 = x, G_2 = \phi^{-1}(y));$$

that is, if $\sigma = \phi$ then $\sigma(G_2) = y$ iff $G_2 = \phi^{-1}(y)$.

Note: In the sequel, we shall use ϕ and τ to denote deterministic elements of S_n , and σ to denote an S_n -valued random variable. Further, for $\phi \in S_n$ we define

$$s(\phi) := \{i \in [n] \text{ s.t. } \phi(i) \neq i\}$$

to be the number of vertices shuffled by the permutation $\phi \in S_n$, and we define

$$S_n(k) =: \{\phi \in S_n \text{ s.t. } s(\phi) = k\}.$$

3 Information loss and graph matching

If the latent alignment between G_1 and G_2 is unknown or errorfully known, graph matching methods can be applied to approximately recover the true alignment. Recalling that $\Pi(n)$ is the set of $n \times n$ permutation matrices, we define the Graph Matching Problem (GMP) as follows:

Definition 5. *Given two graphs G_1 and G_2 in \mathcal{G}_n , with respective adjacency matrices A and B in $\mathbb{R}^{n \times n}$, the graph matching problem is*

$$\min_{P \in \Pi(n)} \|AP - PB\|_F.$$

The GMP objective function $\|AP - PB\|_F$ is equal to $\|A - PBP^T\|_F$ (as the Frobenius norm is invariant under a unitary transformation). In light of this, solving the GMP is equivalent to relabeling the vertices of G_2 so as to minimize the number of induced edge disagreements between A and B . While solving the graph matching problem is NP-hard in general, there are a bevy of approximation algorithms and heuristics in the literature that perform well in practice [55, 49, 16, 14, 28] (in addition, see the excellent survey papers [11, 17] for a thorough review of the prescient literature and discussion of numerous alternate formulations of the GMP). Note that in Section 4, to approximately match the shuffled graphs in our synthetic and real data applications, we use FAQ algorithm of [49] and, when *seeded* vertices are present, the SGM algorithm of [16].

Let $(G_1, G_2) \sim \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$ with respective adjacency matrices A and B . As previously noted, the correlation structure across G_1 and G_2 highlights the natural alignment, id_n , between the two graphs. It is natural to ask for what values of ρ can graph matching recover the latent alignment in the presence of vertex shuffling. To this end, we define

Definition 6. *Let $(G_1, G_2) \sim \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$ with respective adjacency matrices A and B . We say that G_1 and G_2 are matchable if $\arg\min_{P \in \Pi(n)} \|AP - PB\|_F = \{I_n\}$.*

In the correlation regime where G_1 and G_2 are matchable with high probability, what is the degradation in information due to the uncertainty introduced by a random permutation σ ? According to the information processing inequality,

$$I(G_1; G_2) \geq I(G_1; \sigma(G_2)) \tag{3}$$

with equality if and only if σ is a point mass distribution. However, if graph matching can successfully “unshuffle” the graphs—i.e., there is enough signal even in the shuffled graphs to recover the latent alignment—we prove that relatively little information will be lost in the shuffle. We formalize this in the following two theorems, whose proofs can be found in Section A.2 and Section A.3 respectively.

Theorem 7. *Let $(G_1, G_2) \sim \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$, with K and Λ fixed in n , and let σ be uniformly distributed on S_n .*

i. For all values of ρ , it holds that $I(G_1; G_2) - I(G_1; \sigma(G_2)) = O(n \log n)$.

ii. If $\rho = \omega(\sqrt{\log n/n})$ and $\min_i n_i = \Theta(n)$, then $I(G_1; G_2) - I(G_1; \sigma(G_2)) = \Omega(n\rho^2)$.

Note that if ρ is constant in n , then the asymptotic upper bound of Theorem 7 part i. and the asymptotic lower bound of Theorem 7 part ii. differ only by a logarithmic factor. We suspect that the true order is $n \log n$, coinciding with the entropy of the uniformly random shuffling σ .

Remark 8. Theorem 7 part i. is proven with σ uniformly distributed on S_n . We suspect that an analogous result holds for other distributions on S_n that place suitable mass on permutations that shuffle $k = \Theta(n)$ elements of $[n]$, though we do not pursue this further here.

Remark 9. In the proof of Theorem 7 part ii., we essentially prove a stronger statement than that presented in the theorem. If we define $S_n^* := \{\phi \in S_n \mid b(i) = b(\phi(i)) \text{ for all } i \in [n]\}$, to be the set of permutations that preserve vertex block assignment and let σ^* be uniformly distributed on S_n^* , then we prove that under the assumptions of the theorem, $I(G_1; G_2) - I(G_1; \sigma^*(G_2)) = \Omega(n\rho)$. The information processing inequality (see Proposition 25) then gives us that $I(G_1; \sigma(G_2)) \leq I(G_1; \sigma^*(G_2))$; indeed, there is information in the vertices block assignments which is lost in σ and not in σ^* .

In light of Proposition 3, relatively little information is lost due to shuffling in the $\rho = \omega(\sqrt{\log n/n})$ regime: indeed, under this assumption on ρ we have that

$$\frac{I(G_1; G_2) - I(G_1; \sigma(G_2))}{I(G_1; G_2)} = o(1).$$

However, as we will see in Section 4, this information that is lost can have a dramatic negative impact on subsequent inference. We view this relatively minor loss in information as a consequence of ability of graph matching to asymptotically recover the latent alignment almost surely in this correlation regime; see Theorem 10 below.

In the Erdős-Rényi setting ($K = 1$), we proved in [28] that there is a constant $\alpha > 0$ such that if $\rho > \sqrt{\alpha \log n/n}$ then $\operatorname{argmin}_{P \in \Pi} \|A - PBP^T\|_F^2 = \{I\}$ a.a.s. As SBM random graphs are locally Erdős-Rényi, we expect an analogous result to hold in the present SBM setting as well. Indeed, we arrive at the following theorem.

Theorem 10. *With notation as above, let A and B be the adjacency matrices of ρ -SBM(K, \vec{n}, b, Λ) random graphs with K , and Λ fixed in n . For $\tau \in S_n$, define*

$$X_{\tau, A, B} := \|A - P_\tau B P_\tau^T\|_F^2 - \|A - B\|_F^2.$$

There exists a constant $\alpha > 0$ such that if $\rho \geq \sqrt{\alpha \log n/n}$, then

$$\mathbb{P}(\exists \tau \in S_n \text{ with } X_{\tau, A, B} \leq -1) = O(e^{-3 \log n}).$$

The proof of Theorem 10 can be found in Section A.3.

Remark 11. We note here that results similar to Theorem 10 for a much-simplified 2-block SBM appear in [33], although the authors there consider a different MAP-based objective function in their matching.

3.1 The low correlation regime

Let $(G_1, G_2) \sim \rho$ -SBM(K, \vec{n}, b, Λ) with respective adjacency matrices A and B . Theorem 10 asserts that, under mild model assumptions, there exists a constant $\alpha > 0$ such that if $\rho \geq \sqrt{\alpha \log n/n}$ then

$$\lim_{n \rightarrow \infty} \mathbb{P}(\exists \tau \in S_n \text{ with } X_{\tau, A, B} \leq -1) = 0.$$

In Theorem 12 below, we identify a *matchability* phase transition at $\rho = \Theta(\sqrt{\log n/n})$. Indeed, a consequence of Theorem 12 is that, under mild assumptions, there exists a constant $\beta > 0$ such that if $\rho \leq \sqrt{\beta \log n/n}$ then

$$\lim_{n \rightarrow \infty} \mathbb{P}(\exists \tau \in S_n \text{ with } X_{\tau, A, B} \leq -1) = 1.$$

Theorem 12. *With notation as above, let A and B be the adjacency matrices of ρ -SBM(K, \vec{n}, b, Λ) random graphs with K , and Λ fixed in n . Further assume there is an $\eta > 0$ such that $\Lambda \in [\eta, 1 - \eta]^{K \times K}$. Let $\{\tau_i\}_{i=1}^N$ be a collection of $N := \sum_i \lfloor \frac{n_i}{2} \rfloor$ disjoint within-block transpositions; i.e., if $\tau_i = k \leftrightarrow \ell$, then $b(k) = b(\ell) = b(\tau_i(k)) = b(\tau_i(\ell))$. For each $i = 1, 2, \dots, N$, let $E_{\tau_i, A, B}$ be the event $\{X_{\tau_i, A, B} \leq -1\}$. Let $X = \sum_i \mathbb{1}\{E_{\tau_i, A, B}\}$. There exists a constant $\beta > 0$ such that if $\rho \leq \sqrt{\beta \log n/n}$, then $\lim_{n \rightarrow \infty} \mathbb{P}(X = 0) = 0$.*

We conjecture a similar phase transition for the relative information loss due to σ . Indeed, when G_1 and G_2 are not matchable, we conjecture that a nontrivial fraction of the mutual information is lost in the shuffle.

Conjecture 13. *Let $(G_1, G_2) \sim \rho$ -SBM(K, \vec{n}, b, Λ), with K and Λ fixed in n , and let σ be uniformly distributed on S_n . If $\rho = o(\sqrt{\log n/n})$, then*

$$\frac{I(G_1; G_2) - I(G_1; \sigma(G_2))}{I(G_1; G_2)} = \Theta(1).$$

While the matchability phase transition in Theorems 10 and 12 is tighter than the conjectured phase transition in Conjecture 13, if true, Conjecture 13 would imply a duality between information loss and matchability:

1. if $\rho = \omega(\sqrt{\log n/n})$ then $\frac{I(G_1; G_2) - I(G_1; \sigma(G_2))}{I(G_1; G_2)} = o(1)$, and G_1 and G_2 are matchable
2. if $\rho = o(\sqrt{\log n/n})$ then $\frac{I(G_1; G_2) - I(G_1; \sigma(G_2))}{I(G_1; G_2)} = \Theta(1)$, and G_1 and G_2 are not matchable.

We explore the first half of this duality further in the next section. In the $\rho = \omega(\sqrt{\log n/n})$ regime, where the graphs are matchable and relatively little information is lost due to shuffling, we show that graph matching can effectively recover the information lost due to shuffling a.a.s.; see Theorem 15. This provides a theoretical foundation for understanding the utility of graph matching as a pre-processing step for a host of inference tasks: often when the lost information due to shuffling has a negative effect on inference, graph matching can recover the lost information and improve the performance in subsequent inference. In the $\rho = o(\sqrt{\log n/n})$ regime, while the graphs are no longer matchable, we conjecture (and experiments bear out) that the alignment found by graph matching still recovers much of the lost information. However, theoretically working in this regime will require new proof techniques, and we do not pursue this further here.

3.2 Graph matching: Recovering the lost information

In Section 4 we show that the information lost, even when relatively small (see Theorem 7), can have a deleterious effect on subsequent inference, and we demonstrate the potential of graph

matching to recover the lost inferential performance. If $(G_1, \sigma(G_2)) \sim \sigma, \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$, how can we define the matching of $\sigma(G_2)$ to G_1 ?

For $x \in \mathcal{G}_n$ (resp., $y \in \mathcal{G}_n$) with adjacency matrix A (resp., B), we define

$$P_{x,y}^* := \operatorname{argmin}_{P \in P(n)} \|AP - PB\|_F.$$

If $(G_1, G_2) = (x, y)$, then it is natural to define $G_{G_2 \rightarrow G_1}$, G_2 matched to G_1 , as any element of

$$P_{x,y}^*(y) := \{z \in \mathcal{G}_n \text{ s.t. } z = \phi(y) \text{ for } \phi \text{ satisfying } P_\phi \in P_{x,y}^*\},$$

with all elements of $P_{x,y}^*(y)$ being equally probable. Formally, we define

Definition 14. Let $(G_1, G_2) \in \mathcal{G}_n \times \mathcal{G}_n$ be ρ -correlated $\text{SBM}(K, \vec{n}, b, \Lambda)$ graphs. The $\mathcal{G}_n \times \mathcal{G}_n$ -valued random variable $(G_1, G_{G_2 \rightarrow G_1})$ is defined via

$$\begin{aligned} \mathbb{P}[(G_1, G_{G_2 \rightarrow G_1}) = (x, z)] &= \sum_{y \text{ s.t. } z \in P_{x,y}^*(y)} \frac{\mathbb{P}[(G_1, G_2) = (x, y)]}{|P_{x,y}^*(y)|} \\ &= \sum_{y \text{ s.t. } z \in P_{x,y}^*(y)} \frac{\mathbb{P}[(G_1, G_2) = (x, y)]}{|P_{x,z}^*(z)|}. \end{aligned} \quad (4)$$

Moreover, the joint distribution of $(G_1, G_2, G_{G_2 \rightarrow G_1})$ is defined via

$$\begin{aligned} \mathbb{P}[(G_1, G_2, G_{G_2 \rightarrow G_1}) = (x, y, z)] &= \mathbb{P}[(G_1, G_2) = (x, y)] \frac{\mathbb{1}\{z \in P_{x,y}^*(y)\}}{|P_{x,y}^*(y)|} \\ &= \mathbb{P}[(G_1, G_2) = (x, y)] \frac{\mathbb{1}\{z \in P_{x,y}^*(y)\}}{|P_{x,z}^*(z)|}. \end{aligned} \quad (5)$$

In the definition, we made use of the fact that if $z \in P_{x,y}^*(y)$ then $P_{x,z}^*(z) = P_{x,y}^*(y)$. This has two more immediate consequences. First, the normalizing constant $1/|P_{x,z}^*(z)|$ appearing in (4) is precisely what is needed to make $\sum_{x,z} \mathbb{P}[(G_1, G_{G_2 \rightarrow G_1}) = (x, z)] = 1$. Secondly, if we have that $(G_1, \sigma(G_2)) \sim \sigma, \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$, then it is immediate that

$$\mathbb{P}[(G_1, G_{\sigma(G_2) \rightarrow G_1}) = (x, z)] = \mathbb{P}[(G_1, G_{G_2 \rightarrow G_1}) = (x, z)],$$

and

$$\mathbb{P}[(G_1, G_2, G_{\sigma(G_2) \rightarrow G_1}) = (x, y, z)] = \mathbb{P}[(G_1, G_2, G_{G_2 \rightarrow G_1}) = (x, y, z)].$$

In the $\rho = \omega(\sqrt{\log n/n})$ regime, unshuffling $(G_1, \sigma(G_2))$ via graph matching recovers asymptotically almost all of the lost information. While Theorem 7 implies that

$$I(G_1; G_2) - I(G_1; \sigma(G_2)) = \Omega(n\rho),$$

we show below in Theorem 15 that under mild assumptions

$$I(G_1; G_2) - I(G_1; G_{\sigma(G_2) \rightarrow G_1}) = o(1).$$

Theorem 15. *Let $(G_1, \sigma(G_2)) \in \mathcal{G}_n \times \mathcal{G}_n \sim \sigma, \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$ with Λ and K fixed in n . If $\rho = \omega(\sqrt{\log n/n})$, then*

$$I(G_1; G_2) - I(G_1; G_{\sigma(G_2) \rightarrow G_1}) = o(1).$$

The proof of Theorem 15 can be found in Section A.5.

The information processing inequality states that, given random variables X and Y and measurable T ,

$$I(X; Y) \geq I(X; T(Y)).$$

Intuitively, we cannot transform Y independently of X and increase the mutual information between Y and X . At first glance, Theorem 15, which implies that $I(G_1; \sigma(G_2)) < I(G_1; G_{\sigma(G_2) \rightarrow G_1})$, seems to contradict this. However, we note that $G_{\sigma(G_2) \rightarrow G_1}$ is a function of *both* G_1 and $\sigma(G_2)$. Indeed, if $Z = T(X, Y)$ then the information processing inequality need not hold (for a simple example, let X have nontrivial entropy, let Y be independent of X , and let $T(x, y) = x$).

4 The effect on subsequent inference

While the loss in information due to shuffling has little effect on inference tasks that are independent of vertex labels (for example, the nonparametric hypothesis testing methodologies of [44, 4]), the effect on inference that assumes an a priori known vertex alignment may be dramatic. We demonstrate this in the context of joint graph clustering and two sample hypothesis testing for graphs. The trend we demonstrate below is as follows: diminished performance as the alignment is shuffled and the performance loss due to shuffling being recovered via graph matching. We expect this trend to generalize to a host of other joint inference tasks as well. We note here that these results provides a striking contrast to those in [48], where it was shown that even if the graph labeling contain relevant class signal, obfuscating the labels does not necessarily decrease classification performance in a single graph setting.

4.1 Hypothesis Testing

In [45], we propose a semiparametric graph hypothesis testing framework for determining whether two random dot product graphs (RDPG) are generated from the same underlying model.

Definition 16 (d -dimensional RDPG [32, 54]). *Define $\mathcal{X}_d = \{U \in \mathbb{R}^{n \times d} \text{ s.t. } UU^T \in [0, 1]^{n \times n}\}$, and let $X = [X_1 | X_2 | \dots | X_n]^T \in \mathcal{X}_d$. We say that $G_1 \sim \text{RDPG}(X)$ is an instance of a random dot product graph if the adjacency matrix A of G_1 satisfies*

$$\mathbb{P}[A|X] = \prod_{i>j} (X_i^\top X_j)^{A_{ij}} (1 - X_i^\top X_j)^{1-A_{ij}}.$$

In the RDPG model, each vertex v_i in G has a corresponding latent position vector X_i , and given X , G is distributed as a heterogeneous Erdős-Rényi random graph with edge probability matrix $P = XX^T$. If $G \sim \text{SBM}(K, \vec{n}, b, \Lambda)$ with positive semidefinite $\Lambda \in \mathbb{R}^{K \times K}$, then decomposing $\Lambda = \xi \xi^T$ with $\xi = [\xi_1 | \xi_2 | \dots | \xi_K]^T \in \mathbb{R}^{K \times d}$, it is immediate that

$$G \stackrel{\mathcal{L}}{=} G' \sim \text{RDPG}\left(\underbrace{[\xi_1 | \xi_1 | \dots | \xi_1]}_{n_1 \text{ columns}} \underbrace{[\xi_2 | \xi_2 | \dots | \xi_2]}_{n_2 \text{ columns}} \dots \underbrace{[\xi_K | \xi_K | \dots | \xi_K]}_{n_K \text{ columns}}\right)^T.$$

Given two d -dimensional, n -vertex graphs $G_1 \sim \text{RDPG}(X)$ and $G_2 \sim \text{RDPG}(\xi)$ on the same vertex set with the vertex correspondence across graphs known *a priori*, one hypothesis test considered in [45] tests

$$H_0 : Y \perp X \text{ against } H_1 : Y \not\perp X, \quad (6)$$

where $Y \perp X$ if there exists an orthogonal matrix $W \in \mathbb{R}^{d \times d}$ such that $Y = XW$. This formulation accounts for the non-identifiability inherent to RDPG's: if $G_1 \sim \text{RDPG}(X)$ and $G_2 \sim \text{RDPG}(Y)$ with $Y \perp X$, then $G_1 \stackrel{\mathcal{L}}{=} G_2$. If A and B are the respective adjacency matrices of G_1 and G_2 , the test proceeds by first estimating X and Y via the *adjacency spectral embedding* of A and B . To this end, we define:

Definition 17. Let $G \sim \text{RDPG}(X)$ be a d -dimensional, n -vertex RDPG with adjacency matrix A . The adjacency spectral embedding (ASE) of A into \mathbb{R}^d is given by $\hat{X} = U_A S_A^{1/2}$, where

$$|A| = [U_A | \tilde{U}_A] [S_A \oplus \tilde{S}_A] [U_A | \tilde{U}_A]$$

is the spectral decomposition of $|A| = (A^T A)^{1/2}$, $S_A \in \mathbb{R}^{d \times d}$ is the diagonal matrix containing the d largest eigenvalues of $|A|$ on its diagonal, and $U_A \in \mathbb{R}^{n \times d}$ is the matrix whose columns are the corresponding orthonormal eigenvectors.

Under some mild assumptions on the sparsity of $G \sim \text{RDPG}(X)$, in [45] it is proven that

$$\min_{W \in \mathbb{R}^{d \times d} \text{ s.t. } W^T W = I_d} \|\hat{X}W - X\|_F = \Theta(1)$$

with high probability. This fact is leveraged to produce a consistent hypothesis test for testing (6) based on a suitably scaled version of the test statistic

$$T(\hat{X}, \hat{Y}) = \min_{W \in \mathbb{R}^{d \times d} \text{ s.t. } W^T W = I_d} \|\hat{X}W - \hat{Y}\|_F. \quad (7)$$

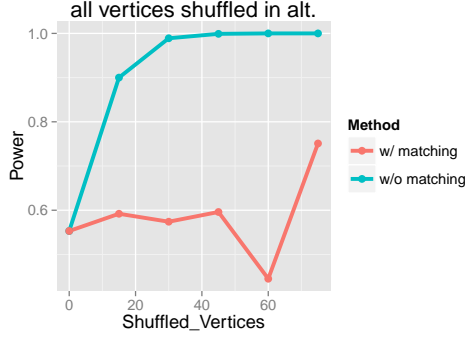
To study the effect that vertex shuffling has in this hypothesis testing framework, we define $X \perp_\ell Y$ if there exists an orthogonal matrix $W \in \mathbb{R}^{d \times d}$ and a permutation matrix P_ϕ , with associated permutation $\phi \in S_n$ satisfying $s(\phi) \leq \ell$, such that $X = P_\phi YW$. In the presence of κ seeded vertices, i.e., vertices whose explicit across graph vertex correspondence is known a priori, we seek here to test the hypotheses

$$H_0^{(n-\kappa)} : Y \perp_{n-\kappa} X \text{ against } H_1^{(n-\kappa)} : Y \not\perp_{n-\kappa} X. \quad (8)$$

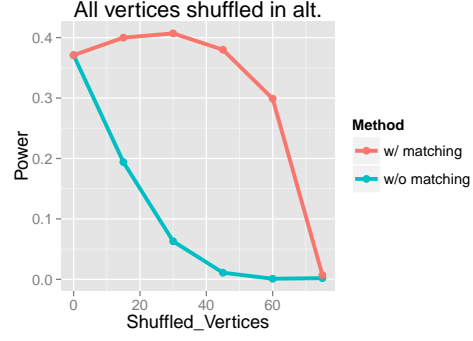
The form of the null hypothesis here arises because, while κ vertices are seeded and are known to be correctly matched across graphs, the veracity of the correspondence between the remaining $n - \kappa$ vertices is unknown a priori. Note that $\kappa = 0$ corresponds to the case $H_0^{(n)}$, where the vertices are shuffled by some permutation in S_n .

To illustrate the effect of shuffling in the present testing paradigm, we set

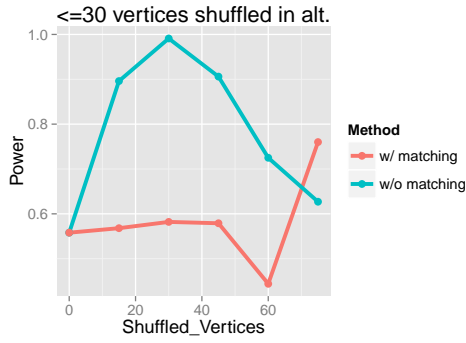
$$\Lambda_1 = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.3 & 0.5 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}, \quad \Lambda_2 = \begin{bmatrix} 0.8 & 0.3 & 0.2 \\ 0.3 & 0.8 & 0.3 \\ 0.2 & 0.3 & 0.8 \end{bmatrix}, \quad \Lambda_3 = \begin{bmatrix} 0.5 & 0.4 & 0.3 \\ 0.4 & 0.5 & 0.4 \\ 0.3 & 0.4 & 0.5 \end{bmatrix},$$



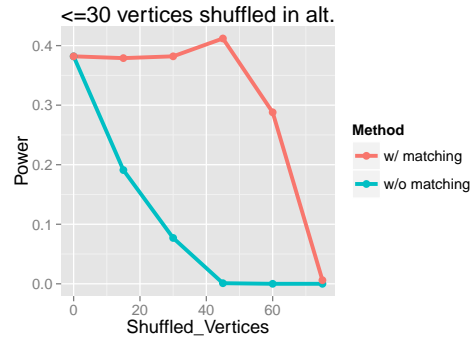
(a) All unseeded vertices shuffled in G_2 under H_1



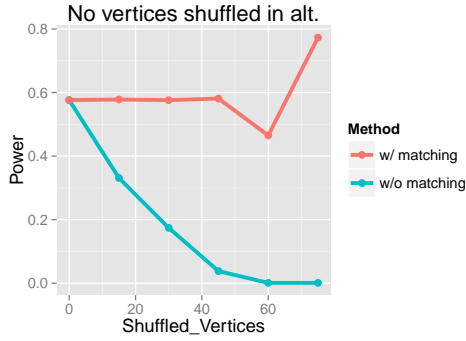
(b) All unseeded vertices shuffled in G_3 under H_1



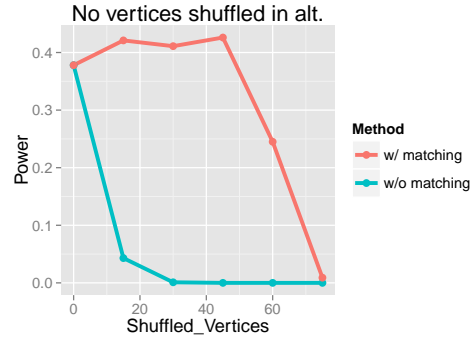
(c) At most 30 vertices shuffled in G_2 under H_1



(d) At most 30 vertices shuffled in G_3 under H_1



(e) No vertices shuffled in G_2 under H_1



(f) No vertices shuffled in G_3 under H_1

Figure 1: Let G_1 , G_2 and G_3 be as in (9) and (10). In the left (resp., right) column, we plot the power of the above procedure for testing (where we view X as fixed and vary Y in $H_1^{(k)}$) $H_0^{(k)} : Y_2 \not\cong_k X$ (resp., $H_0^{(k)} : Y_3 \not\cong_k X$) versus $H_1^{(k)} : Y_2 \cong_k X$ (resp., $H_1^{(k)} : Y_3 \cong_k X$) for various values of k (i.e., for various levels of shuffling under the null hypothesis). Each column contains three plots representing whether the vertex correspondence between G_1 and G_2 (left column) or G_1 and G_3 (right column) under $H_1^{(k)}$ satisfies the following: in (a), (b) all unseeded vertices are shuffled in the alternate; in (c), (d) the unseeded vertices are partially shuffled in the alternate (min(30, 75 - k) unseeded vertices shuffled); and in (e), (f) the unseeded vertices are all unshuffled in the alternate. The blue curves represent the power of the above test for the hypotheses in (8) while the red curves represent first using the seeded vertices to approximately match the two graphs via the SGM algorithm [16] and then testing the hypotheses in (6). In each case, the power computations are based on 1000 Monte Carlo trials.

noting that each Λ_i is positive semidefinite. Let (G_1, G_2) be 0.4-correlated SBMs satisfying

$$G_1 \sim \text{SBM}(3, [25, 25, 25], b, \Lambda_1), \quad G_2 \sim \text{SBM}(3, [25, 25, 25], b, \Lambda_2). \quad (9)$$

In addition, let (G_1, G_3) be 0.4-correlated SBMs satisfying

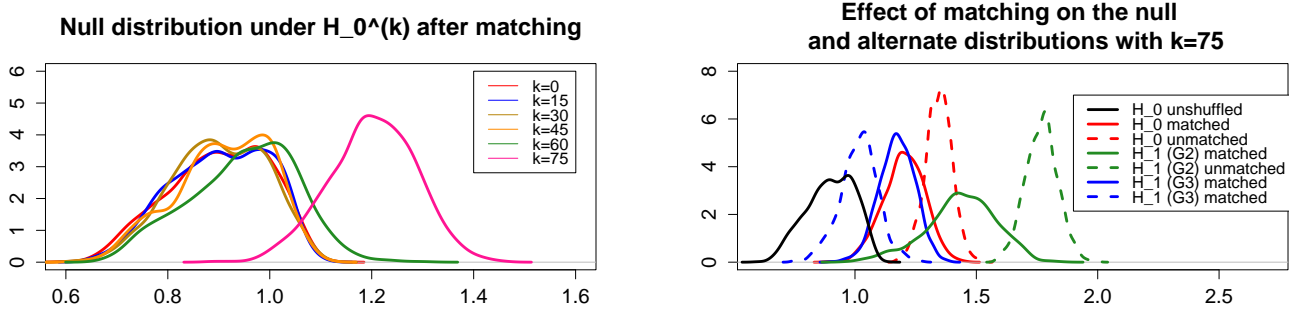
$$G_1 \sim \text{SBM}(3, [25, 25, 25], b, \Lambda_1), \quad G_3 \sim \text{SBM}(3, [25, 25, 25], b, \Lambda_3). \quad (10)$$

In all cases, $b(i) = 1$ if $1 \leq i \leq 25$, $b(i) = 2$ if $26 \leq i \leq 50$, and $b(i) = 3$ if $51 \leq i \leq 75$. Note that the definition of correlated SBM's with different Λ 's is completely analogous to the equal Λ case of Definition 1, and so is omitted. As each Λ is positive semi-definite, we can realize G_1 (resp., G_2 and G_3) as $G_1 \sim \text{RDPG}(X)$ (resp., $G_2 \sim \text{RDPG}(Y_2)$ and $G_3 \sim \text{RDPG}(Y_3)$) for appropriately chosen X (resp., Y_2 and Y_3). The null distribution can be sampled from in this example, and so we empirically compute the level-0.05 critical value of our test based on the statistic T of (7) without reliance on the bootstrapping procedure of [45, Algorithm 1]. Note that the null distribution should be computed under the least favorable element of the null hypothesis; i.e., $n - \kappa$ vertices shuffled. Note that across all simulations, we use the true $d = 3$ as the embedding dimension.

In the left (resp., right) column of Figure 1, we plot the power of the above procedure for testing (where we view X as fixed and vary Y in $H_1^{(k)}$) $H_0^{(k)} : Y_2 \perp_k X$ (resp., $H_0^{(k)} : Y_3 \perp_k X$) versus $H_1^{(k)} : Y_2 \not\perp_k X$ (resp., $H_1^{(k)} : Y_3 \not\perp_k X$) for various values of k (i.e., for various levels of shuffling in the null). Each column contains three plots representing whether the vertex correspondence between G_1 and G_2 (resp., G_3 in the right column) under $H_1^{(k)}$ satisfies the following: in (a), (b) all unseeded vertices are shuffled in the alternate; in (c), (d) the unseeded vertices are partially shuffled in the alternate ($\min(30, 75 - k)$ unseeded vertices shuffled); and in (e), (f) the unseeded vertices are all unshuffled in the alternate. The blue curves represent the power of the above test for the hypotheses in (8) while the red curves represent first using the seeded vertices to approximately match the two graphs via the SGM algorithm [16]—to ameliorate the effect of shuffling—and then directly testing the hypotheses in (6).

Unlike the uniformly negative effect on clustering performance we will see later in Section 4.2, the effect that vertex shuffling has on the power of testing (8) is more nuanced. In Figure 1 panels (a) and (c) we see that for some instantiations of $H_1^{(k)}$ —demonstrated via $G_2 \sim \text{RDPG}(Y_2)$ here—shuffling can improve testing power without the need for graph matching. In these cases, if the correspondence between G_1 and G_2 under the alternate distribution is sufficiently shuffled, then the power of our procedure for testing (8) is increased *after shuffling*, and the matched test has diminished power when compared to the shuffled test. However, if the correspondence between G_1 and G_2 under the alternate distribution is sufficiently unshuffled compared to the correspondence under the null distribution, we see that the power of our procedure for testing (8) is greatly diminished; see (e) in Figure 1. Indeed, information is lost between (G_1, G_2) under the null distribution due to progressively more shuffling, and that same information is not lost in the alternate. Graph matching recovers the lost signal and ameliorates the effect of shuffling on testing power. As, in practice, the exact level of shuffling amongst the unseeded vertices is unknown a priori, we propose the graph matching version of the test as a conservative, more robust, version of testing (8).

For other instantiations of $H_1^{(k)}$ —demonstrated via $G_3 \sim \text{RDPG}(Y_3)$ here—shuffling dramatically decreases testing power. In these cases—see (b), (d), and (f) in Figure 1—the loss of information between G_1 and G_3 due to shuffling has a significant negative effect on testing power



(a) Distribution of the test statistic T under $H_0^{(k)}$ after graph matching for various values of k

(b) Effect of GM on the null and alternate distributions when $k = 75$

Figure 2: In the left panel, we plot the distribution of the test statistic T under $H_0^{(k)}$ after graph matching for various values of k . The bias introduced into the distribution of T for $k = 75$ is a consequence of the failure of our GM algorithm to recover the true correspondence without any seeds. In the Right panel, we plot the effect of GM on the null (in red) and alternate distributions (with (G_1, G_2) in green; (G_1, G_3) in blue). Note the different effects that shuffling has on the alternate distribution: for (G_1, G_2) shuffling introduces helpful bias which is partially mitigated via GM; for (G_1, G_3) shuffling introduces harmful bias which is not able to be fixed via the GM algorithm we employ. In each case, the density estimates are based on 1000 Monte Carlo trials.

and the more vertices shuffled the worse the power of the test. Moreover, as long as $k < 75$ in $H_0^{(k)}$, graph matching recovers the lost information and subsequently the lost testing power. We view the decreased power at $k = 75$ in panels (b), (d) and (f) of Figure 1 as an algorithmic artifact. Indeed, with no seeds the GM algorithm we employ often does not recover the true correspondence after shuffling, and the bias introduced into the null distribution of T by the incorrect matching is the reason for the decreased power; see Figure 2. With a perfect matching, we would expect power to be approximately 0.4 in these cases—i.e., in panels (b), (d) and (f)—across all k . This also accounts for the sharp power increase in the matched test in panels (a), (c) and (e) of Figure 1 at $k = 75$. The bias introduced in this case leads to increased testing power which is not fully corrected (unlike in the seeded case) via matching, and this bias has a beneficial effect on testing power. The increased variance after matching is due to the presence of multiple local minima in the search space of the GM algorithm. With a perfect matching, we would expect power to be approximately 0.6 in these cases—i.e., in panels (a), (c) and (e)—across all k .

While the matched test has marginally decreased power at some elements of $H_1^{(k)}$, the shuffled test has dramatically decreased power (compared to the matched test) at other elements of $H_1^{(k)}$. We posit that the matched test has nontrivial power over a broader range of the alternative than the unmatched (shuffled) test, and we again propose the graph matching version of the test as a conservative, more robust, version of testing (8). However, we are not presently able to classify which elements of the alternative lead to a relatively more powerful test before or after matching, and we are working to understand which elements of $H_1^{(k)}$ achieve higher power under each test.

Remark 18. In practice, to adapt the test in [45] to (8), we can modify the bootstrapping procedure outlined in [45, Algorithm 1] to compute the p -value of the test in (8) based on the test

Algorithm 1 Bootstrapping procedure for the test $H_0^{(k)}: Y \perp_k X$ versus $H_1^{(k)}: Y \not\perp_k X$.

```

1: procedure BOOTSTRAP( $X, T, k, bs, \rho$ )
2:    $d = \text{ncol}(X)$  ▷  $d$  is the number of columns in  $X$ 
3:    $\mathcal{S}_X = \emptyset$ 
4:   for  $b = 1 : bs$  do
5:      $(A_b, B_b) = \rho\text{-correlated RDPG}(X)$ ; ▷  $\rho$ -correlated heterogeneous ER( $XX^\top$ ) graphs
6:      $\phi = \text{Unif}(S_n(k))$ ;  $P = P_\phi$  ▷ Select  $\phi$  uniformly from  $S_n(k)$ 
7:      $\hat{X}_b = \text{ASE}(A_b, d)$ ;  $\hat{Y}_b = \text{ASE}(B_b, d)$  ▷ ASE  $A_b$  and  $B_b$  into  $\mathbb{R}^{n \times d}$ 
8:      $\hat{Y}_b = P\hat{Y}_b$  ▷ Permute the rows of  $Y$  via  $P$ 
9:      $T_b = \min_{W \in \mathbb{R}^{d \times d_{\text{s.t.}}}, W^\top W = I_d} \|\hat{X}_b W - \hat{Y}_b\|_F$  ▷ Bootstrapped test statistic  $T_b$ 
10:     $\mathcal{S}_X = \mathcal{S}_X \cup T_b$ 
11:   end for
12:   return  $p = (|\{s \in \mathcal{S}_X : s \geq T\}| + 0.5)/bs$  ▷ Returns the p-value (with continuity correction) associated with  $T$ 
13:
14: end procedure
15:
16:  $\hat{X} = \text{ASE}(A, d)$ ;  $\hat{Y} = \text{ASE}(B, d)$  ▷  $d$  is known a priori
17:  $T = \min_{W \in \mathbb{R}^{d \times d_{\text{s.t.}}}, W^\top W = I_d} \|\hat{X} W - \hat{Y}\|_F$ 
18:  $p_X = \text{Bootstrap}(\hat{X}, T, k, bs, \rho)$  ▷  $bs$ , the number of bootstrap samples, is known a priori;  $\rho$  is known a priori
19:
20:  $p_Y = \text{Bootstrap}(\hat{Y}, T, k, bs, \rho)$ 
21:  $p = \max\{p_X, p_Y\}$  ▷ Return the maximum of the two p-values.

```

statistic (7). Note that the p -value should be computed under the least favorable element of the null hypothesis; i.e., $n - \kappa$ vertices shuffled. The modified bootstrapping procedure is summarized in Algorithm 1. Note that the inclusion of a correlation factor ρ into Algorithm 1 is done to connect the testing work to our present ρ -correlated regime, as the procedure in [45] does not specify (nor exclude the presence of) such a parameter.

4.2 Joint versus single graph clustering

We next explore the impact that label shuffling has on spectral graph clustering. Spectral graph clustering has become an important and widely-used machine learning method, with a sizable literature devoted to various spectral clustering algorithms under several model assumptions; see, for example, [25, 36, 39, 43, 15, 31]. We focus here on a variant of the methodology of [43], which embeds the adjacency matrix of a graph into an appropriate Euclidean space and subsequently employs the k -means algorithm to cluster the data. Here, rather than using k -means clustering to cluster the data, we will employ the model-based clustering algorithm **Mclust** [18].

Formally, given a suitable embedding dimension d and adjacency matrix A (see Remark 19 for comments on how to practically select d), Adjacency Spectral Clustering (ASC) proceeds by first computing the d -dimensional ASE \hat{X} of A , and then clustering the rows of \hat{X} into k clusters using k -means or the model-based clustering algorithm **Mclust**. In addition to being easily and efficiently implemented on very large graphs, adjacency spectral clustering lends itself to tractable

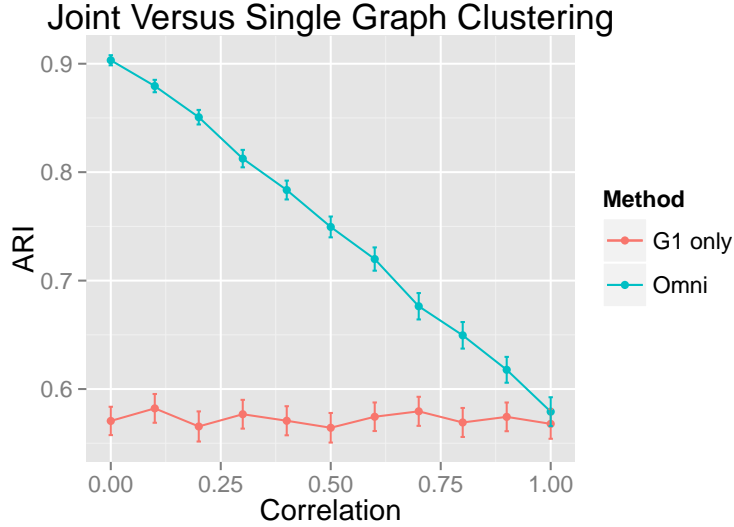


Figure 3: Joint versus single graph clustering of two ρ -correlated SBM. Over a range of ρ , we embed G_1 and G_2 via joint ASC and in blue plot the mean Adjusted Rand Index (ARI) ± 2 s.e. for the clustering of G_1 against its true block assignments. We also apply ASC to G_1 alone, and in red we plot the mean ARI ± 2 s.e. against the true block assignments for single graph clustering. In each case the number of Monte Carlo trials was 1000. Note that, except in the case of very high correlation, there is a significant performance increase achieved when clustering G_1 after jointly embedding G_1 and G_2 versus embedding G_1 alone.

theoretical analysis. For example, in [43, 15] the k -means variant of ASC is proven to, with high probability, consistently recover the block assignments in an SBM under mild model assumptions; and in [31] the k -means variant of ASC is proven to, with high probability, perfectly recover the block assignments in an SBM under mildly more stringent model assumptions. Due to its solid theoretical foundation, and excellent performance in practice (see, for example, [43, 15]), ASC has been a popular and well studied algorithm in the network literature.

Remark 19 (Model Selection). In practice, we often need to estimate both d and k in the ASC procedure. While this model selection problem is theoretically daunting [15], as a useful heuristic method we can use singular value thresholding [8] to estimate d from a partial SCREE plot of the eigenvalues of $|A|$. To estimate k , we can use traditional measures of clustering validity such as silhouette width or a Bayesian information criterion (BIC) penalization.

When we have multiple graph valued observations of the same data, can we efficiently utilize the information between the graphs to increase clustering performance? In the manifold matching literature, there are numerous examples of this heuristic: leveraging the signal across multiple data sets can increase inference performance within each of the data sets (see, for example, [35, 29, 42, 40]). Inspired by this, in [9] the authors adapted the methodology of [35, 29] for joint graph inference and showed the potential for significantly increased classification performance by jointly classifying even modestly correlated graphs. Adapting their methodology to joint graph clustering, we proceed as outlined in Algorithm 2

To demonstrate the potential performance increase achievable via joint ASC versus single graph

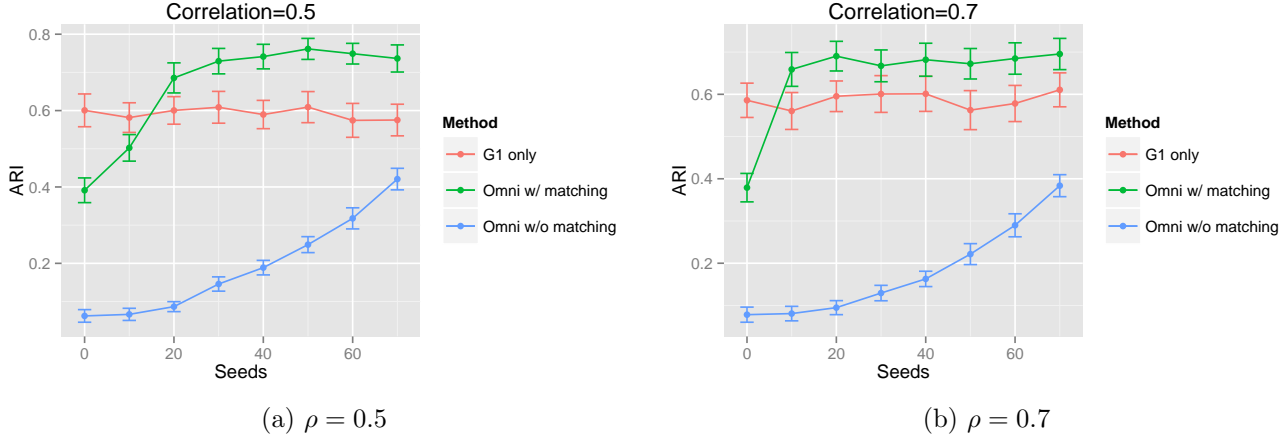


Figure 4: Joint clustering performance with errorfully observed labeling across graphs. In blue, we plot the mean ARI ± 2 s.e. of the clustering of G_1 obtained via Joint ASC (against the true block assignments of G_1) when $100 - s$ of the vertices in G_2 have their labels randomly permuted. In red, we plot the mean ARI ± 2 s.e. of ASC applied only to G_1 (against the true block assignments of G_1), and in green we plot the mean ARI ± 2 s.e. of the clustering of G_1 obtained via Joint ASC (against the true block assignments of G_1) after matching the shuffled B back to A . In the left plot, the across graph correlation is 0.5, and is 0.7 in the rightmost plot. Note that the performance of joint clustering dramatically decreases as more labels in the graph are shuffled, though this performance loss is recoverable via graph matching.

ASC, we consider Joint ASC on

$$(G_1, G_2) \sim \rho - \text{SBM} \left(2, \vec{n}, b, \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.2 \end{bmatrix} \right),$$

with $n_i = 50$ for each $i = 1, 2$, $b(i) = 1$ if $1 \leq i \leq 50$, and $b(i) = 2$ if $51 \leq i \leq 100$. Results are displayed in Figure 3. Over a range of ρ , we embed G_1 and G_2 via Joint ASC and in blue plot the mean Adjusted Rand Index [37] (ARI) ± 2 s.e. for the clustering obtained for G_1 against its true block assignments. We also apply ASC to G_1 alone, and in red we plot the mean ARI ± 2 s.e. of the obtained clustering of G_1 against its true block assignments. In each case the number of Monte Carlo trials was 1000. Across these synthetic experiments, we used the true parameters $d = k = 2$.

Algorithm 2 Joint ASC

Input: Matched adjacency matrices $A \in \{0, 1\}^{n \times n}$ and $B \in \{0, 1\}^{n \times n}$; embedding dimension d .

Output: Joint clustering of graphs into k clusters

Step 1: Create the Omnibus adjacency matrix

$$\mathcal{O} = \begin{bmatrix} A & \frac{A+B}{2} \\ \frac{A+B}{2} & B \end{bmatrix}$$

Step 2: Perform ASC on the omnibus matrix \mathcal{O}

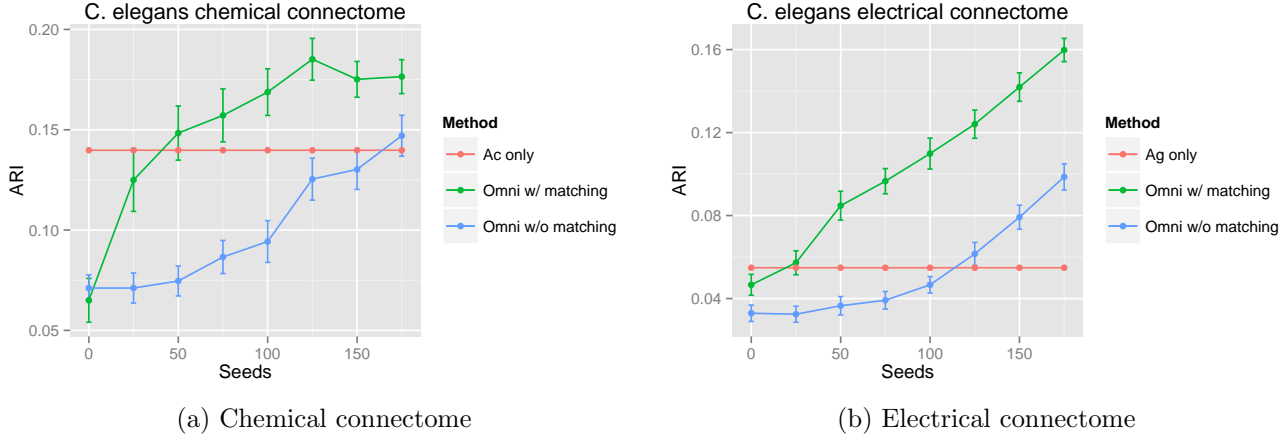


Figure 5: Joint clustering performance on the *C. elegans* connectome with errorfully observed labeling. In blue, we plot the mean ARI ± 2 s.e. (against the true clustering of the connectome into inter, motor, and sensory neurons) of Joint ASC when $247 - s$ of the vertices in G_2 (chosen uniformly at random) have their labels randomly permuted. In red, we plot the mean ARI ± 2 s.e. of ASC applied only to the single graph, and in green we plot the mean ARI ± 2 s.e. of Joint ASC after matching the shuffled B back to A . In the left plot, we plot the performance on the chemical connectome, and on the right the electrical connectome. Note that the performance of joint clustering dramatically decreases as more labels in the graph are shuffled, though this performance loss is recoverable via graph matching.

In Figure 3, we see significantly improved clustering accuracy achieved by joint inference for modest to lowly correlated (G_1, G_2). Note that as the correlation increases, the increased performance due to the borrowed strength of joint inference diminishes. This is unsurprising as the amount of additional information added by G_2 is less for larger ρ (indeed if $\rho = 0$ then $H(G_2|G_1) = H(G_2) = O(n^2)$ while if $\rho = 1$ then $H(G_2|G_1) = 0$). Does the increased performance due to joint inference degrade in the presence of an errorfully observed vertex correspondence? To explore this further, we randomly permute the labels of $100 - s$ vertices in B , so that there are $s \in \{0, 10, 20, 30, 40, 50, 60, 70\}$ seeded vertices whose labels are kept true (note that these seeded vertices are randomly chosen from the 100 total vertices). We plot the performance of Joint ASC pre and post graph matching for $\rho = 0.5$ and $\rho = 0.7$ in Figure 4. In light of Theorem 7, we see that the modest information lost due to the shuffling dramatically decreases the performance of Joint ASC. This can be readily explained: the information lost due to shuffling is precisely the information leveraged by joint ASC, namely the block assignments of the vertices across graphs.

Remark 20. Note that, in the presence of errorfully known or unknown vertex correspondences across graphs, we can not embed $(A + B)/2$ directly. Indeed, in the presence of labeling errors, we would not be able to evaluate the clustering performance (as we can when running Algorithm 2) when clustering the embedded points of $(A + B)/2$.

In Figure 4, we also explore the inferential impact of graph matching in recovering this lost performance. In light of Theorem 15, we see that graph matching recovers the information lost in shuffling, and therefore recovers much of lost performance. It is notable that when $s = 0$,

GM using FAQ performs little better than chance in recovering the true across graph labeling (recall, exact GM is NP-hard and note that seeding was shown in [16] to dramatically increase GM performance). While this was also seen in [16, 28], we note that FAQ does (errorfully) recover the across-graph block assignments, and so aligns the graphs in a way that preserves some of the necessary structure leveraged by Joint ASC. This suggests an extension of Theorems 15, in which perfect matching is not needed for a significant portion of the information lost by shuffling to be fully recovered.

Note also that in the lower correlation setting, the performance after matching is significantly better than the corresponding performance in the higher correlation setting (relative to the base single graph inference level). This confirms the intuition proposed by Figure 3: With higher correlation, there is less additional information to be recovered by matching (although fewer seeds are needed to recover this lost information), and therefore, there is a comparatively smaller increase in performance achieved by the joint inference even post matching.

As a second example, we consider jointly clustering the *C. elegans* connectomes [47]. The connectome of the *C. elegans* roundworm has been completely mapped and neurons interact with each other in two distinct modes: via electrical gap junctions and via chemical synapses. In [26, 9], the authors showed that the electrical and chemical connectomes contain complimentary signal, and both papers suggest that inference should proceed in the joint graph space. Here we explore not only what gains are achieved via joint inference, but how these gains are lost via errorful labelings, and whether GM can recover these lost gains. To this end, as in [9], we preprocess the data by removing the isolated neurons (under either modality) and symmetrizing each connectome. The resulting connectomes each have 253 vertices which are classified into 3 neural types: motor neurons, sensory neurons, and inter neurons. In the left (resp., right) panel of Figure 5, we plot the performance of Joint ASC in clustering the chemical (resp., electrical) connectome versus single graph ASC applied only to the chemical (resp., electrical) connectome. Throughout, we estimate d as in Remark 19 and use `Mclust` to cluster the data into $k = 3$ clusters.

In both modalities, significantly better clustering performance is achieved by working in the joint graph space; indeed, in the case where the correspondence is perfectly observed across graphs, the ARI for the clustering of the chemical (resp., electrical) connectome in Joint ASC is 0.24 (resp., 0.22) compared to 0.14 (resp., 0.05) ARI via single graph ASC. Also, as we observed in the synthetic data, this gain is lost when a significant portion of the graphs have the across graph correspondence shuffled (when the number of seeded vertices is ≤ 100 in the chemical connectome and ≤ 130 in the electrical connectome). These graphs are particularly challenging to match, with only $\approx 10\%$ of the vertices correctly matched by the state-of-the-art SGM algorithm even with 150 seeds (see [16, Figure 4]). Nonetheless, the structure that is uncovered by graph matching (namely the recovery of vertex classes across graphs) is enough to recover a large portion of the performance increase seen in perfectly matched Joint ASC. Again, this suggests an extension of Theorems 15, in which perfect matching is not needed for a significant portion of the information lost by shuffling to be fully recovered.

5 Discussion and future work

Many joint graph inference procedures assume that the vertex correspondence between graphs is known a priori. However, in practice the correspondence may only be partially known or errorfully known, and we seek to understand the effect that errors in the labeling have on subsequent

inference. To this end, we provide an information theoretical foundation for answering the following questions: What is the increase in uncertainty (i.e., loss of the mutual information) between two graphs when the labeling across graphs is errorfully observed, and can this lost information be recovered via graph matching? Working in the correlated stochastic blockmodel setting, we prove that when graph matching can perfectly recover an errorfully observed correspondence (Theorem 10), relatively little information is lost due to shuffling (Theorem 7). However, we demonstrate that this lost information can have a dramatic effect on the performance of subsequent inference. We also show that asymptotically almost all of the lost information can be recovered via graph matching (Theorem 15), which has the effect of recovering much of the lost inferential performance.

In the process, we are able to establish a phase transition for graph matchability at $\rho = \Theta(\sqrt{\log n/n})$ for $(G_1, G_2) \sim \rho$ -SBM. We prove in Theorems 10 and 12 that under mild assumptions there exists constants $0 < \beta < \alpha$ such that $\rho \geq \sqrt{\alpha \log n/n}$ implies G_1 and G_2 are matchable and $\rho \leq \sqrt{\beta \log n/n}$ implies G_1 and G_2 are unmatchable. We conjecture the analogous phase transition at $\rho = \Theta(\sqrt{\log n/n})$ for the relative information loss due to shuffling in Conjecture 13. Establishing Conjecture 13 would cement a duality between the information lost due to shuffling and matchability (i.e., the ability to undo the shuffling via graph matching): if $\rho = \omega(\sqrt{\log n/n})$ then $\frac{I(G_1; G_2) - I(G_1; \sigma(G_2))}{I(G_1; G_2)} = o(1)$, and G_1 and G_2 are matchable; if $\rho = o(\sqrt{\log n/n})$ then $\frac{I(G_1; G_2) - I(G_1; \sigma(G_2))}{I(G_1; G_2)} = \Theta(1)$, and G_1 and G_2 are not matchable. The difficulty in proving the conjecture lies in lower bounding the mutual information in the mixture model $(G_1, \sigma(G_2))$ with low correlation.

While graph matching cannot correctly recover the lost correspondence in the low correlation setting, $\rho = o(\sqrt{\log n/n})$, we suspect—and the experiments of Section 4 demonstrate—that a significant portion of the lost information is recovered even by an imperfect matching. This would have at least two immediate consequences. First, the estimated correlation across different real data networks—even on the same vertex sets—is often very small. A theorem proving that GM can recover much of the lost information with an imperfect matching in these low correlation regimes would further highlight the applicability of GM across a broad class of data sets. Second, as graph matching is NP-hard in general and no algorithm exists that can perfectly match even modestly sized graphs, extending Theorem 15 to the case of an imperfect matching would serve to further highlight the practical utility of graph matching *algorithms*.

Lastly, can we extend this theory to a broader class of random graph models? While stochastic blockmodels are widely used to model data with latent community structure, they are an overly simplistic model for many real data applications. While extending the theory to ρ -correlated Inhomogeneous ER graphs is the natural next step (which is reasonable in light of [27]), we also wish to extend the theory to non-edge-independent random graph models (for example, power law graphs, bounded degree graphs, etc.). However, these non-edge-independent graphs require a novel correlation structure and new graph matching theory to be developed.

6 Acknowledgments

We wish to thank Donniell Fishkind, Carey Priebe, Daniel Sussman, Minh Tang, Avanti Athreya and Joshua T. Vogelstein for their comments and suggestions which greatly helped the exposition and ideas in this paper.

A Proofs and supporting results

Herein, we collect the proofs of the main theorems and supporting results.

A.1 Proof of Proposition 3

In this section, we will provide a proof of Proposition 3. Recall that the random variables

$$\{\mathbb{1}[\{j, k\} \in E(G_i)]\}_{i=1,2;\{j,k\} \in \binom{V}{2}}$$

are collectively independent except that for each $\{j, k\} \in \binom{V}{2}$, the correlation between $\mathbb{1}[\{j, k\} \in E(G_1)]$ and $\mathbb{1}[\{j, k\} \in E(G_2)]$ is $\rho \geq 0$. Next note that if $(X, Y) \sim \rho$ -correlated $\text{Bern}(p)$ then

$$\begin{aligned} I(X; Y) &= p(p + \rho(1 - p)) \log \left(1 + \rho \frac{(1 - p)}{p} \right) + 2p(1 - p)(1 - \rho) \log(1 - \rho) \\ &\quad + (1 - p)(1 - p + p\rho) \log \left(1 + \rho \frac{p}{1 - p} \right). \end{aligned}$$

Together this yields

$$\begin{aligned} I(G_1; G_2) &= \sum_{g_1, g_2 \in \mathcal{G}_n} \mathbb{P}(G_1 = g_1, G_2 = g_2) \log \left(\frac{\mathbb{P}(G_1 = g_1, G_2 = g_2)}{\mathbb{P}(G_1 = g_1)\mathbb{P}(G_2 = g_2)} \right) \\ &= \sum_{i,j \in [K] \leq j}^K \left[n_{i,j} \Lambda_{i,j} (\Lambda_{i,j} + \rho(1 - \Lambda_{i,j})) \log \left(1 + \rho \frac{(1 - \Lambda_{i,j})}{\Lambda_{i,j}} \right) + 2n_{i,j} \Lambda_{i,j} (1 - \Lambda_{i,j})(1 - \rho) \log(1 - \rho) \right. \\ &\quad \left. + n_{i,j} (1 - \Lambda_{i,j})(1 - \Lambda_{i,j} + \Lambda_{i,j}\rho) \log \left(1 + \rho \frac{\Lambda_{i,j}}{1 - \Lambda_{i,j}} \right) \right], \end{aligned}$$

where

$$n_{i,j} = \begin{cases} n_i n_j & \text{if } i \neq j; \\ \binom{n_i}{2} & \text{if } i = j. \end{cases}$$

Next, note that for any fixed $p \in (0, 1)$, if $\rho \rightarrow 0$ as $n \rightarrow \infty$, then for sufficiently large n

$$\begin{aligned} &p(p + \rho(1 - p)) \log \left(1 + \rho \frac{(1 - p)}{p} \right) + 2p(1 - p)(1 - \rho) \log(1 - \rho) \\ &\quad + (1 - p)(1 - p + p\rho) \log \left(1 + \rho \frac{p}{1 - p} \right) \\ &= (p^2 + \rho p(1 - p)) \left(\rho \frac{(1 - p)}{p} - \rho^2 \frac{(1 - p)^2}{2p^2} + O(\rho^3) \right) + 2p(1 - p)(1 - \rho) \left(-\rho - \frac{\rho^2}{2} + O(\rho^3) \right) \\ &\quad + (1 - p)(1 - p + p\rho) \left(\rho \frac{p}{1 - p} - \rho^2 \frac{p^2}{2(1 - p)^2} + O(\rho^3) \right) \\ &= \frac{\rho^2}{2} + O(\rho^3). \end{aligned}$$

The proof follows immediately.

A.2 Proof of Theorem 7

Under the assumptions of the theorem, if σ is uniformly distributed on S_n , then for any $\tau \in S_n$,

$$\mathbb{P}(\sigma(G_2) = \tau(g_2)) = \sum_{\phi \in S_n} \frac{1}{|S_n|} \mathbb{P}(G_2 = \phi \circ \tau(g_2)) = \sum_{\phi \in S_n} \frac{1}{|S_n|} \mathbb{P}(G_2 = \phi(g_2)) = \mathbb{P}(\sigma(G_2) = g_2). \quad (11)$$

We also have that

$$\begin{aligned} I(G_1; \sigma(G_2)) &= \sum_{\substack{g_1 \in \mathcal{G}_n, \\ g_2 \in \mathcal{G}_n}} \left(\sum_{\phi \in S_n} \frac{1}{|S_n|} \mathbb{P}(G_1 = g_1, G_2 = \phi(g_2)) \right) \log \left(\sum_{\tau \in S_n} \frac{1}{|S_n|} \frac{\mathbb{P}(G_1 = g_1, G_2 = \tau(g_2))}{\mathbb{P}(G_1 = g_1) \mathbb{P}(\sigma(G_2) = g_2)} \right) \\ &= \sum_{\phi \in S_n} \frac{1}{|S_n|} \sum_{\substack{g_1 \in \mathcal{G}_n, \\ g_2 \in \mathcal{G}_n}} \mathbb{P}(G_1 = g_1, G_2 = g_2) \log \left(\sum_{\tau \in S_n} \frac{1}{|S_n|} \frac{\mathbb{P}(G_1 = g_1, G_2 = \tau \circ \phi^{-1}(g_2))}{\mathbb{P}(G_1 = g_1) \mathbb{P}(\sigma(G_2) = \phi^{-1}(g_2))} \right), \end{aligned}$$

so that applying Eq. (11) yields

$$\begin{aligned} &I(G_1; G_2) - I(G_1; \sigma(G_2)) \\ &= \sum_{\phi \in S_n} \frac{1}{|S_n|} \sum_{\substack{g_1 \in \mathcal{G}_n, \\ g_2 \in \mathcal{G}_n}} \mathbb{P}(G_1 = g_1, G_2 = g_2) \log \left(\frac{\mathbb{P}(G_1 = g_1, G_2 = g_2) \mathbb{P}(\sigma(G_2) = g_2)}{\sum_{\tau \in S_n} \frac{1}{|S_n|} \mathbb{P}(G_1 = g_1, G_2 = \tau \circ \phi^{-1}(g_2)) \mathbb{P}(G_2 = g_2)} \right) \\ &= - \sum_{\phi \in S_n} \frac{1}{|S_n|} \sum_{\substack{g_1 \in \mathcal{G}_n, \\ g_2 \in \mathcal{G}_n}} \mathbb{P}(G_1 = g_1, G_2 = g_2) \log \left(\sum_{\tau \in S_n} \frac{1}{|S_n|} \frac{\mathbb{P}(G_1 = g_1, G_2 = \tau \circ \phi^{-1}(g_2))}{\mathbb{P}(G_1 = g_1, G_2 = g_2)} \right) \\ &\quad + \sum_{g_2 \in \mathcal{G}_n} \mathbb{P}(G_2 = g_2) \log \left(\frac{\mathbb{P}(\sigma(G_2) = g_2)}{\mathbb{P}(G_2 = g_2)} \right). \end{aligned} \quad (12)$$

Proof of Theorem 7 part i. We begin by proving part i. of the theorem. Note that

$$\sum_{g_2 \in \mathcal{G}_n} \mathbb{P}(G_2 = g_2) \log \left(\frac{\mathbb{P}(\sigma(G_2) = g_2)}{\mathbb{P}(G_2 = g_2)} \right) = -H(\sigma(G_2)) + H(G_2). \quad (13)$$

Concavity of the entropy function $H(\cdot)$ then yields

$$-H(\sigma(G_2)) + H(G_2) \leq - \sum_{\phi \in S_n} \frac{1}{|S_n|} H(\phi(G_2)) + H(G_2) = 0.$$

Applying this to Eq. (12) then yields

$$\begin{aligned} &I(G_1; G_2) - I(G_1; \sigma(G_2)) \\ &\leq - \sum_{\phi \in S_n} \frac{1}{|S_n|} \sum_{\substack{g_1 \in \mathcal{G}_n, \\ g_2 \in \mathcal{G}_n}} \mathbb{P}(G_1 = g_1, G_2 = g_2) \log \left(\sum_{\tau \in S_n} \frac{1}{|S_n|} \frac{\mathbb{P}(G_1 = g_1, G_2 = \tau \circ \phi^{-1}(g_2))}{\mathbb{P}(G_1 = g_1, G_2 = g_2)} \right) \\ &\leq - \sum_{\phi \in S_n} \frac{1}{|S_n|} \sum_{\substack{g_1 \in \mathcal{G}_n, \\ g_2 \in \mathcal{G}_n}} \mathbb{P}(G_1 = g_1, G_2 = g_2) \log \left(\frac{1}{|S_n|} \right) = - \log \left(\frac{1}{|S_n|} \right) = \log(|S_n|) \sim (n \log n), \end{aligned}$$

as desired. \square

Proof of Theorem 7 part ii. Let $(G_1, G_2) \sim \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$. To prove part ii. of Theorem 7, we will consider permutations in

$$S_n^* := \{\phi \in S_n \mid b(i) = \phi(b(i)) \text{ for all } i \in [n]\},$$

i.e., permutations of the vertex sets of G_1 and G_2 that fix block assignments. For σ^* uniformly distributed in S_n^* , we will show that

$$I(G_1; G_2) - I(G_1; \sigma^*(G_2)) = \Omega(n^2 \rho^2).$$

To complete the proof, we use the information processing inequality to show that $I(G_1; \sigma(G_2)) \leq I(G_1; \sigma^*(G_2))$ (see Proposition 25 for detail).

We now establish some notation and preliminary results. For $\phi \in S_n^*$, write $\phi = (\phi_1, \phi_2, \dots, \phi_K)$, where $\phi_i : V_i \mapsto V_i$ is the restriction of ϕ to V_i .

Definition 21. Let (G_1, G_2) be ρ -correlated SBM(K, \vec{n}, b, Λ) random graphs, and let $x, y \in \mathcal{G}_n$.

1. For each of $i = 1, 2$, and $j = 1, 2, \dots, K$, let $G_i^j = G_i|_{V_j}$ (resp., $x^j = x|_{V_j}$, $y^j = y|_{V_j}$) be the induced subgraph of G_i (resp., fixed $x, y \in \mathcal{G}_n$) restricted to V_j .
2. For $j, \ell \in [K]$ with $j < \ell$, let $G_i^{j,\ell}$ (resp., $x^{j,\ell}$, $y^{j,\ell}$) be the induced bipartite subgraph of G_i (resp., x, y) composed of the edges in G_i (resp., x, y) between vertex sets V_j and V_ℓ .

The key to restricting our attention to permutations in S_n^* is the following. Working with permutations in S_n^* allows us to split the SBM random graphs along block assignments, and then tackle each block (and each bipartite between-block) subgraph separately. Before formalizing this in Claim 23, we will need to define the analogues of correlated bipartite graphs. To this end, if \mathcal{B}_{m_1, m_2} is the set of all labeled bipartite graphs $G = (U, V, E)$ with $|U| = m_1$, $|V| = m_2$, and $E \subset U \times V$, then we define:

Definition 22. Two $m_1 m_2$ -vertex bipartite random graphs $(G_1, G_2) \in \mathcal{B}_{m_1, m_2} \times \mathcal{B}_{m_1, m_2}$ are ρ -correlated Bipartite(m_1, m_2, p) random graphs (abbreviated ρ -Bipartite) if

- i. For each $i = 1, 2$, $G_i \in \mathcal{B}_{m_1, m_2}$, and edges between the bipartite sets U and V are independently present with common probability p ;
- ii. The random variables $\{\mathbb{1}[(j, k) \in E(G_i)]\}_{i=1,2; j \in U, k \in V}$ are collectively independent except that for each $(j, k) \in U \times V$, the correlation between $\mathbb{1}[(j, k) \in E(G_1)]$ and $\mathbb{1}[(j, k) \in E(G_2)]$ is $\rho \geq 0$.

The deterministic shuffling of ρ -Bipartite graphs can be defined completely analogously to Eq. (2). To wit, if $x \in \mathcal{B}_{m_1, m_2}$, and $\tau \in S_{m_1}$, $\phi \in S_{m_2}$, we define the $[\tau, \phi]$ -shuffled graph, denoted by $[\tau, \phi](x) = (V, E_{[\tau, \phi](x)}) \in \mathcal{B}_{m_1, m_2}$, via

$$(i, j) \in E_x \text{ iff } (\tau(i), \phi(j)) \in E_{[\tau, \phi](x)};$$

i.e., U is shuffled according to τ and V is shuffled according to ϕ . The following claim is immediate.

Claim 23. Let $(G_1, G_2) \sim \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$, and let $\phi \in S_n^*$.

1. $\mathbb{P}(G_2 = y) = \mathbb{P}(G_2 = \phi(y))$ for all $y \in \mathcal{G}_n$;
2. For each $j = 1, 2, \dots, K$, (G_1^j, G_2^j) are distributed as ρ -ER($n_j, B_{j,j}$) random graphs;
3. For $j, \ell \in [K]$ with $j < \ell$, $(G_1^{j,\ell}, G_2^{j,\ell})$ are distributed as ρ -Bipartite($n_j, n_\ell, B_{j,\ell}$);
4. The collection of graph pairs

$$\{(G_1^i, G_2^i)\}_{i \in [K]} \bigcup \{(G_1^{j,\ell}, G_2^{j,\ell})\}_{j,\ell \in [K], j < \ell}$$

is mutually independent.

Analogues of Theorem 10 hold in the ρ -ER and ρ -Bipartite settings as well. The following Lemma is proved similarly to Theorem 10, and so the proof is only briefly sketched.

Lemma 24. *With notation as above,*

- i) If $(G_1, G_2) \sim \rho$ -ER(m, p) with respective adjacency matrices A and B , and if $\rho = \omega\left(\sqrt{\frac{\log m}{m}}\right)$ then there exists a $C > 0$ such that

$$\mathbb{P}(\exists \phi \in S_n \text{ s.t. } \|A - P_\phi B P_\phi^T\|_F^2 - \|A - B\|_F^2 \leq C m \rho) = O(e^{-4 \log m}). \quad (14)$$

- ii) If $(G_1, G_2) \sim \rho$ -Bipartite(m_1, m_2, p) with respective adjacency matrices A and B , and if $\rho = \omega\left(\sqrt{\frac{\log(m_1+m_2)}{m_1+m_2}}\right)$ then there exists a $C > 0$ such that

$$\begin{aligned} \mathbb{P}(\exists (\phi, \tau) \in S_{m_1} \times S_{m_2} \text{ s.t. } \|A - P_\phi B P_\tau^T\|_F^2 - \|A - B\|_F^2 \leq C(m_1 \wedge m_2)\rho) \\ = O(e^{-4 \log(m_1+m_2)}), \end{aligned} \quad (15)$$

where $m_1 \wedge m_2 = \min(m_1, m_2)$.

Proof. We will sketch the proof of part i) with part ii) following mutadis mutandis. For the moment, fix $\phi \in S_m$ with $s(\phi) = k$. For $x, y \in \mathcal{G}_n$, define

$$\begin{aligned} F_{\mathcal{A}}(x, y, \phi) &:= \left\{ \{u, v\} \in \binom{V}{2} \text{ s.t. } u \approx_x v, u \sim_{\phi(x)} v, \text{ and } u \approx_{\phi(y)} v \right\}; \\ F_{\mathcal{O}}(x, y, \phi) &:= \left\{ \{u, v\} \in \binom{V}{2} \text{ s.t. } u \sim_x v, u \approx_{\phi(x)} v, \text{ and } u \sim_{\phi(y)} v \right\}. \end{aligned}$$

We then have (if x and y have adjacency matrices A and B)

$$\frac{1}{2}(\|A - P_\phi B P_\phi^T\|_F^2 - \|A - B\|_F^2) = \frac{1}{2}\|A - P_\phi A P_\phi^T\|_F^2 - 2F_{\mathcal{A}}(x, y, \phi) - 2F_{\mathcal{O}}(x, y, \phi).$$

With \mathbf{m}_k defined via $\mathbf{m}_k = \binom{k}{2} + k(m-k)$, applying [23, Proposition 3.2] to $(G_1, G_2) \sim \rho$ -ER(m, p) yields that there exists a constant c such that for m sufficiently large

$$\mathbb{P}\left(G = g \text{ s.t. } \left| \frac{1}{2}\|A_g - P_\phi A_g P_\phi^T\|_F^2 - 2\mathbf{m}_k p(1-p) \right| \geq 2c\sqrt{\mathbf{m}_k k \log m} \right) \leq 2e^{-3k \log m}. \quad (16)$$

Conditioning on $\frac{1}{2}\|A - P_\phi A P_\phi^T\|_F^2 = \Delta$, $F_{\mathcal{O}}(G_1, G_2, \phi) \sim \text{Bin}(\Delta/2, p(1 - \rho))$ independent of $F_{\mathcal{A}}(G_1, G_2, \phi) \sim \text{Bin}(\Delta/2, (1 - p)(1 - \rho))$. Simple applications of Hoeffding's inequality then yield that

$$\mathbb{P}\left[F_{\mathcal{O}}(G_1, G_2, \phi) \geq \frac{\Delta}{2} \left(p(1 - \rho) + \frac{\rho}{3}\right) \middle| \frac{1}{2}\|A - P_\phi A P_\phi^T\|_F^2 = \Delta\right] \leq e^{-\Delta\rho^2/9}; \quad (17)$$

$$\mathbb{P}\left[F_{\mathcal{A}}(G_1, G_2, \phi) \geq \frac{\Delta}{2} \left((1 - p)(1 - \rho) + \frac{\rho}{3}\right) \middle| \frac{1}{2}\|A - P_\phi A P_\phi^T\|_F^2 = \Delta\right] \leq e^{-\Delta\rho^2/9}. \quad (18)$$

Unconditioning (17)–(18) combined with (16), and summing over k yields the desired result. \square

The utility of Eq. (14) and (15) in the present ρ -SBM setting can be realized as follows. For each $i, j \in [K]$, we define

$$\xi_{i,j} := 1 + \frac{\rho}{(1 - \Lambda_{i,j})\Lambda_{i,j}(1 - \rho)^2} > 1.$$

For ease of notation, we will write $\mathbb{P}(g_1, g_2) := \mathbb{P}(G_1 = g_1, G_2 = g_2)$ for all $g_1, g_2 \in \mathcal{G}_n$. Combining the above yields

$$\begin{aligned} I(G_1; G_2) - I(G_1; \sigma^*(G_2)) &= \sum_{\substack{g_1 \in \mathcal{G}_n, \\ g_2 \in \mathcal{G}_n}} \mathbb{P}(g_1, g_2) \log \left(\frac{\mathbb{P}(g_1, g_2)}{\sum_{\tau \in S_n^*} \frac{1}{|S_n^*|} \mathbb{P}(g_1, \tau(g_2))} \right) \\ &= - \sum_{\substack{g_1 \in \mathcal{G}_n, \\ g_2 \in \mathcal{G}_n}} \mathbb{P}(g_1, g_2) \log \left(\sum_{\tau \in S_n^*} \frac{1}{|S_n^*|} \frac{\mathbb{P}(g_1, \tau(g_2))}{\mathbb{P}(g_1, g_2)} \right) \\ &= - \sum_{\substack{g_1 \in \mathcal{G}_n, \\ g_2 \in \mathcal{G}_n}} \mathbb{P}(g_1, g_2) \log \left(\sum_{\tau \in S_n^*} \frac{1}{|S_n^*|} \prod_{j \in [K]} \exp \left\{ \log(\xi_{j,j}) \frac{1}{4} \left(\|A \upharpoonright_{g_1^j} - B \upharpoonright_{g_2^j}\|_F^2 - \|A \upharpoonright_{g_1^j} - P_{\tau_j} B \upharpoonright_{g_2^j} P_{\tau_j}^T\|_F^2 \right) \right\} \right. \\ &\quad \left. \prod_{\substack{j, \ell \in [K] \\ \text{s.t. } j < \ell}} \exp \left\{ \log(\xi_{j,\ell}) \frac{1}{4} \left(\|A \upharpoonright_{g_1^{j,\ell}} - B \upharpoonright_{g_2^{j,\ell}}\|_F^2 - \|A \upharpoonright_{g_1^{j,\ell}} - P_{\tau_j} B \upharpoonright_{g_2^{j,\ell}} P_{\tau_\ell}^T\|_F^2 \right) \right\} \right) \end{aligned}$$

If $\min_i n_i = \Theta(n)$ and $\rho = \omega\left(\sqrt{\frac{\log n}{n}}\right)$, for n sufficiently large there exists constants $C > 0$, $C' > 0$ such that for all $\tau \in S_n^*$,

$$-\frac{1}{4} \left(\|A \upharpoonright_{g_1^{j,\ell}} - P_{\tau_j} B \upharpoonright_{g_2^{j,\ell}} P_{\tau_\ell}^T\|_F^2 - \|A \upharpoonright_{g_1^{j,\ell}} - B \upharpoonright_{g_2^{j,\ell}}\|_F^2 \right) < -Cn\rho$$

and

$$-\frac{1}{4} \left(\|A \upharpoonright_{g_1^j} - P_{\tau_j} B \upharpoonright_{g_2^j} P_{\tau_j}^T\|_F^2 - \|A \upharpoonright_{g_1^j} - B \upharpoonright_{g_2^j}\|_F^2 \right) < -Cn\rho$$

with probability at least $1 - C'e^{-4 \log n}$. Therefore, for n sufficiently large there exists a constant $C'' > 0$ such that

$$I(G_1; G_2) - I(G_1; \sigma^*(G_2)) \geq C''n\rho^2 + C'e^{-4 \log n}n^2 = \Omega(n\rho^2) \quad (19)$$

as desired. The proof is then completed by applying the information processing inequality to show that $I(G_1; \sigma(G_2)) \leq I(G_1; \sigma^*(G_2))$. To this end, we have the following proposition.

Proposition 25. *Let σ be uniformly distributed on S_n independent of σ^* uniformly distributed on S_n^* . With notation as above, $I(G_1; \sigma(G_2)) \leq I(G_1; \sigma^*(G_2))$.*

Proof. For $x, y, z \in \mathcal{G}_n$,

$$\mathbb{P}(G_1 = x, \sigma^*(G_2) = y, \sigma(G_2) = z) = \mathbb{P}(G_1 = x) \mathbb{P}(\sigma^*(G_2) = y, \sigma(G_2) = z | G_1 = x),$$

and

$$\begin{aligned} \mathbb{P}(\sigma^*(G_2) = y, \sigma(G_2) = z | G_1 = x) &= \sum_{\phi \in S_n} \sum_{\phi^* \in S_n^*} \frac{\mathbb{P}(G_2 = \phi^*(y), G_2 = \phi(z) | G_1 = x)}{|S_n| \cdot |S_n^*|} \\ &= \sum_{\phi \in S_n} \sum_{\phi^* \in S_n^*} \frac{\mathbb{P}(G_2 = \phi^*(y) | G_1 = x)}{|S_n^*|} \frac{\mathbb{1}\{\phi(z) = \phi^*(y)\}}{|S_n|} \\ &= \sum_{\phi^* \in S_n^*} \frac{|\{\phi \in S_n \text{ s.t. } \phi(z) = \phi^*(y)\}|}{|S_n|} \frac{\mathbb{P}(G_2 = \phi^*(y) | G_1 = x)}{|S_n^*|} \end{aligned}$$

Lastly, noting that

$$|\{\phi \in S_n \text{ s.t. } \phi(z) = \phi^*(y)\}| = |\{\phi \in S_n \text{ s.t. } \phi(z) = y\}|$$

and

$$\mathbb{P}(\sigma(G_2) = z | \sigma^*(G_2) = y) = \frac{|\{\phi \in S_n \text{ s.t. } \phi(z) = y\}|}{|S_n|},$$

we have that

$$\begin{aligned} \mathbb{P}(G_1 = x, \sigma^*(G_2) = y, \sigma(G_2) = z) \\ = \mathbb{P}(G_1 = x) \mathbb{P}(\sigma^*(G_2) = y | G_1 = x) \mathbb{P}(\sigma(G_2) = z | \sigma^*(G_2) = y) \end{aligned}$$

The information processing inequality then yields that $I(G_1; \sigma(G_2)) \leq I(G_1; \sigma^*(G_2))$. \square

A.3 Proof of Theorem 10

Herein, we prove Theorem 10. We first restate the Theorem.

Theorem 10. *With notation as above, let A and B be the adjacency matrices of ρ -SBM(K, \vec{n}, b, Λ) random graphs with K , and Λ fixed in n . For $\tau \in S_n$, define $X_{\tau, A, B} := \frac{1}{2}(\|A - P_\tau B P_\tau^T\|_F^2 - \|A - B\|_F^2)$. There exists a constant $\alpha > 0$ such that for $\rho \geq \sqrt{\alpha \frac{\log(n)}{n}}$, we have*

$$\mathbb{P}(\exists \tau \in S_n \text{ with } X_{\tau, A, B} \leq -1) = O(e^{-3 \log n}).$$

Proof. Fix $\tau \neq \text{id}_n \in S_n$, and suppose that τ permutes the labels of exactly $m \geq 2$ vertices (so that $|\{v : \tau(v) = v\}| = n - m$). For each pair $1 \leq i, j \leq K$, let

$$\epsilon_{i,j}^\tau := |\{v \in V_i \text{ s.t. } \tau(v) \in V_j, v \neq \tau(v)\}|,$$

and let

$$f_i^\tau = |\{v \in V_i \text{ s.t. } \tau(v) = v\}|.$$

Note that for each $i \in [K]$ and each $\tau \in S_n$, we have that

$$n_i - f_i^\tau = \sum_j \epsilon_{i,j}^\tau = \sum_j \epsilon_{j,i}^\tau.$$

As in the proof of Lemma 24, we note that if $x, y \in \mathcal{G}_n$ with adjacency matrices A_x and B_y respectively,

$$X_{\tau, A_x, B_y} := \frac{1}{2}(\|A_x - P_\tau B_y P_\tau^T\|_F^2 - \|A_x - B_y\|_F^2) = \frac{1}{2}\|A_x - P_\tau A_x P_\tau^T\|_F^2 - 2F_{\mathcal{A}}(x, y, \tau) - 2F_{\mathcal{O}}(x, y, \tau),$$

where

$$F_{\mathcal{A}}(x, y, \tau) := \left\{ \{u, v\} \in \binom{V}{2} \text{ s.t. } u \asymp_x v, u \sim_{\tau(x)} v, u \asymp_{\tau(y)} v \right\}, \text{ and}$$

$$F_{\mathcal{O}}(x, y, \tau) := \left\{ \{u, v\} \in \binom{V}{2} \text{ s.t. } u \sim_x v, u \asymp_{\tau(x)} v, u \sim_{\tau(y)} v \right\}.$$

We call the errors induced by τ on x of the form $u \asymp_x v$ and $u \sim_{\tau(x)} v$ *addition errors* (so that $F_{\mathcal{A}}$ is the number of fixed addition errors), and the errors induced by τ on x of the form $u \sim_x v$ and $u \asymp_{\tau(x)} v$ *occlusion errors* (so that $F_{\mathcal{O}}$ is the number of fixed occlusion errors).

We will first show that if $(G_1, G_2) \sim \rho\text{-SBM}(K, \vec{n}, b, \Lambda)$ satisfying the assumptions in the theorem, then with sufficiently high probability $(G_1, G_2) = (x, y)$ satisfying

$$\frac{1}{2}\|A_x - P_\tau A_x P_\tau^T\|_F^2 > 2F_{\mathcal{A}}(x, y, \tau) + 2F_{\mathcal{O}}(x, y, \tau),$$

implying that $X_{\tau, A_x, B_y} > 0$. To this end, with A and B the random adjacency matrices associated with G_1 and G_2 respectively, note that

$$\begin{aligned} \frac{1}{2}\|A - P_\tau A^T P_\tau^T\|_F^2 &= \sum_{\substack{\{v, v'\} \in \binom{V}{2} \\ \tau(v) \neq v \text{ or } \tau(v') \neq v'}} (A_{v, v'} - A_{\tau(v), \tau(v')})^2 \\ &= \sum_{\substack{\{v, v'\} \in \binom{V}{2} \\ \tau(v) \neq v \text{ or } \tau(v') \neq v'}} A_{v, v'}(1 - A_{\tau(v), \tau(v')}) + (1 - A_{v, v'})A_{\tau(v), \tau(v')} \\ &= \sum_{i=1}^K \sum_{\substack{\{v, v'\} \in \binom{V_i}{2} \\ \tau(v) \neq v \text{ or } \tau(v') \neq v'}} A_{v, v'}(1 - A_{\tau(v), \tau(v')}) + (1 - A_{v, v'})A_{\tau(v), \tau(v')} \end{aligned} \quad (20)$$

$$+ \sum_{i=1}^K \sum_{j>i}^K \sum_{\substack{(v, v') \in V_i \times V_j \\ \tau(v) \neq v \text{ or } \tau(v') \neq v'}} A_{v, v'}(1 - A_{\tau(v), \tau(v')}) + (1 - A_{v, v'})A_{\tau(v), \tau(v')}. \quad (21)$$

Consider the sum in (20). For each $i \in [K]$, the sum can be further decomposed into three terms:

1. For each $j \in [K]$ there are $\mathbf{n}_1(i, i, j, j) := \binom{\epsilon_{i,j}^\tau}{2}$ terms with both $v \neq \tau(v)$ and $v' \neq \tau(v')$ mapped to V_j by τ . The expected number of addition errors (denoted $\mathcal{A}_{i,i,j,j}^{(1)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{A}_{i,i,j,j}^{(1)}) = \binom{\epsilon_{i,j}^\tau}{2} (1 - \Lambda_{i,i}) \Lambda_{j,j},$$

and the expected number of occlusion errors (denoted $\mathcal{O}_{i,i,j,j}^{(1)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{O}_{i,i,j,j}^{(1)}) = \binom{\epsilon_{i,j}^\tau}{2} \Lambda_{i,i} (1 - \Lambda_{j,j}).$$

Conditioning on $\mathcal{A}_{i,i,j,j}^{(1)}$ and $\mathcal{O}_{i,i,j,j}^{(1)}$, each addition error is independently corrected in B with probability $(1 - \rho)(1 - \Lambda_{j,j})$, and each occlusion error is independently corrected (independently also of the corrected addition errors) in B with probability $(1 - \rho)\Lambda_{j,j}$.

2. For each $j \in [K]$, $\ell \in [K]$, $\ell > j$, there are $\mathbf{n}_2(i, i, j, \ell) := \epsilon_{i,j}^\tau \epsilon_{i,\ell}^\tau$ terms with $v \neq \tau(v) \in V_j$ and $v' \neq \tau(v') \in V_\ell$. The expected number of addition errors (denoted $\mathcal{A}_{i,i,j,\ell}^{(2)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{A}_{i,i,j,\ell}^{(2)}) = \epsilon_{i,j}^\tau \epsilon_{i,\ell}^\tau (1 - \Lambda_{i,i}) \Lambda_{j,\ell},$$

and the expected number of occlusion errors (denoted $\mathcal{O}_{i,i,j,\ell}^{(2)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{O}_{i,i,j,\ell}^{(2)}) = \epsilon_{i,j}^\tau \epsilon_{i,\ell}^\tau \Lambda_{i,i} (1 - \Lambda_{j,\ell}).$$

Conditioning on $\mathcal{A}_{i,i,j,\ell}^{(2)}$ and $\mathcal{O}_{i,i,j,\ell}^{(2)}$, each addition error is independently corrected in B with probability $(1 - \rho)(1 - \Lambda_{j,\ell})$, and each occlusion error is independently corrected (independently also of the corrected addition errors) in B with probability $(1 - \rho)\Lambda_{j,\ell}$.

3. For each $j \in [K]$, there are $\mathbf{n}_3(i, i, i, j) := f_i^\tau \epsilon_{i,j}^\tau$ terms with $v = \tau(v) \in V_i$ and $v' \neq \tau(v') \in V_j$. The expected number of addition errors (denoted $\mathcal{A}_{i,i,i,j}^{(3)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{A}_{i,i,i,j}^{(3)}) = f_i^\tau \epsilon_{i,j}^\tau (1 - \Lambda_{i,i}) \Lambda_{i,j},$$

and the expected number of occlusion errors (denoted $\mathcal{O}_{i,i,i,j}^{(3)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{O}_{i,i,i,j}^{(3)}) = f_i^\tau \epsilon_{i,j}^\tau \Lambda_{i,i} (1 - \Lambda_{i,j}).$$

Conditioning on $\mathcal{A}_{i,i,i,j}^{(3)}$ and $\mathcal{O}_{i,i,i,j}^{(3)}$, each addition error is independently corrected in B with probability $(1 - \rho)(1 - \Lambda_{i,j})$, and each occlusion error is independently corrected (independently also of the corrected addition errors) in B with probability $(1 - \rho)\Lambda_{i,j}$.

In the sum in (21), for each $i, j \in [K]$ with $j > i$, the sum can be further decomposed into three terms:

4. For each $\ell \in [K]$, $h \in [K]$, there are $\mathbf{n}_4(i, j, h, \ell) := \epsilon_{i,h}^\tau \epsilon_{j,\ell}^\tau$ terms with $v \in V_i$, $v \neq \tau(v) \in V_h$ and $v' \in V_j$, $v' \neq \tau(v') \in V_\ell$. The expected number of addition errors (denoted $\mathcal{A}_{i,j,h,\ell}^{(4)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{A}_{i,j,h,\ell}^{(4)}) = \epsilon_{i,h}^\tau \epsilon_{j,\ell}^\tau (1 - \Lambda_{i,j}) \Lambda_{h,\ell},$$

and the expected number of occlusion errors (denoted $\mathcal{O}_{i,j,h,\ell}^{(4)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{O}_{i,j,h,\ell}^{(4)}) = \epsilon_{i,h}^\tau \epsilon_{j,\ell}^\tau \Lambda_{i,j} (1 - \Lambda_{h,\ell}).$$

Conditioning on $\mathcal{A}_{i,j,h,\ell}^{(4)}$ and $\mathcal{O}_{i,j,h,\ell}^{(4)}$, each addition error is independently corrected in B with probability $(1 - \rho)(1 - \Lambda_{h,\ell})$, and each occlusion error is independently corrected (independently also of the corrected addition errors) in B with probability $(1 - \rho)\Lambda_{h,\ell}$.

5. For each $\ell \in [K]$, there are $\mathbf{n}_5(i, j, i, \ell) := f_i^\tau \epsilon_{j,\ell}^\tau$ terms with $v = \tau(v) \in V_i$ and $v' \in V_j$, $v' \neq \tau(v') \in V_\ell$. The expected number of addition errors (denoted $\mathcal{A}_{i,j,i,\ell}^{(5)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{A}_{i,j,i,\ell}^{(5)}) = f_i^\tau \epsilon_{j,\ell}^\tau (1 - \Lambda_{i,j}) \Lambda_{i,\ell},$$

and the expected number of occlusion errors (denoted $\mathcal{O}_{i,j,i,\ell}^{(5)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{O}_{i,j,i,\ell}^{(5)}) = f_i^\tau \epsilon_{j,\ell}^\tau \Lambda_{i,j} (1 - \Lambda_{i,\ell}).$$

Conditioning on $\mathcal{A}_{i,j,i,\ell}^{(5)}$ and $\mathcal{O}_{i,j,i,\ell}^{(5)}$, each addition error is independently corrected in B with probability $(1 - \rho)(1 - \Lambda_{i,\ell})$, and each occlusion error is independently corrected (independently also of the corrected addition errors) in B with probability $(1 - \rho)\Lambda_{i,\ell}$.

6. For each $\ell \in [K]$, there are $\mathbf{n}_6(i, j, \ell, j) := f_j^\tau \epsilon_{i,\ell}^\tau$ terms with $v' = \tau(v') \in V_j$ and $v \in V_i$, $v \neq \tau(v) \in V_\ell$. The expected number of addition errors (denoted $\mathcal{A}_{i,j,\ell,j}^{(6)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{A}_{i,j,\ell,j}^{(6)}) = f_j^\tau \epsilon_{i,\ell}^\tau (1 - \Lambda_{i,j}) \Lambda_{\ell,j},$$

and the expected number of occlusion errors (denoted $\mathcal{O}_{i,j,\ell,j}^{(6)}$) contributed by these terms is

$$\mathbb{E}(\mathcal{O}_{i,j,\ell,j}^{(6)}) = f_j^\tau \epsilon_{i,\ell}^\tau \Lambda_{i,j} (1 - \Lambda_{\ell,j}).$$

Conditioning on $\mathcal{A}_{i,j,\ell,j}^{(6)}$ and $\mathcal{O}_{i,j,\ell,j}^{(6)}$, each addition error is independently corrected in B with probability $(1 - \rho)(1 - \Lambda_{\ell,j})$, and each occlusion error is independently corrected (independently also of the corrected addition errors) in B with probability $(1 - \rho)\Lambda_{\ell,j}$.

For each $s \in [6]$ and each feasible set of indices $(a, b, c, d) \in K^4$, note that $\mathbf{n}_s(a, b, c, d) = O(mn)$. If $\mathbf{n}_s(a, b, c, d) \geq m\sqrt{n \log n}$, then an application of [23, Proposition 3.2] yields that there exists a constant $\gamma > 0$ such that for n sufficiently large

$$\mathbb{P} \left(|\mathcal{A}_{a,b,c,d}^{(s)} - \mathbb{E}\mathcal{A}_{a,b,c,d}^{(s)}| > \gamma \sqrt{\mathbf{n}_s(a, b, c, d)} \sqrt{m \log n} \right) \leq 2e^{-3m \log n}, \quad (22)$$

and

$$\mathbb{P} \left(|\mathcal{O}_{a,b,c,d}^{(s)} - \mathbb{E}\mathcal{O}_{a,b,c,d}^{(s)}| > \gamma \sqrt{\mathbf{n}_s(a, b, c, d)} \sqrt{m \log n} \right) \leq 2e^{-3m \log n}. \quad (23)$$

Alternatively, if $\mathbf{n}_s(a, b, c, d) < m\sqrt{n \log n}$, then it is immediate that

$$|\mathcal{A}_{a,b,c,d}^{(s)} - \mathbb{E}\mathcal{A}_{a,b,c,d}^{(s)}| \leq m\sqrt{n \log n}, \text{ and } |\mathcal{O}_{a,b,c,d}^{(s)} - \mathbb{E}\mathcal{O}_{a,b,c,d}^{(s)}| \leq m\sqrt{n \log n}. \quad (24)$$

Denote by \mathcal{E} the event that A satisfies

$$|\mathcal{A}_{a,b,c,d}^{(s)} - \mathbb{E}\mathcal{A}_{a,b,c,d}^{(s)}| \leq \gamma\sqrt{\mathbf{n}_s(a,b,c,d)}\sqrt{m\log n}, \quad |\mathcal{O}_{a,b,c,d}^{(s)} - \mathbb{E}\mathcal{O}_{a,b,c,d}^{(s)}| \leq \gamma\sqrt{\mathbf{n}_s(a,b,c,d)}\sqrt{m\log n}$$

for all $s \in [6]$ and feasible indices $(a,b,c,d) \in K^4$ satisfying $\mathbf{n}_s(a,b,c,d) \geq m\sqrt{n\log n}$. Eq. (22)–(23) and a simple union bound imply that

$$\mathbb{P}(\mathcal{E}^c) \leq 24K^4 e^{-3m\log n}. \quad (25)$$

For each $s \in [6]$ and each feasible set of indices $(a,b,c,d) \in K^4$, consider constants

$$x_{a,b,c,d}^{(s)}, y_{a,b,c,d}^{(s)} \in [0, \gamma\sqrt{\mathbf{n}_s(a,b,c,d)}\sqrt{m\log n}], \text{ and } w_{a,b,c,d}^{(s)}, z_{a,b,c,d}^{(s)} \in [0, m\sqrt{n\log n}].$$

Let $\mathcal{E}_{\mathbf{x},\mathbf{y},\mathbf{w},\mathbf{z}} \subset \mathcal{E}$ be the event that for all $s \in [6]$ and each feasible set of indices $(a,b,c,d) \in K^4$, we have

$$\begin{aligned} |\mathcal{A}_{a,b,c,d}^{(s)} - \mathbb{E}\mathcal{A}_{a,b,c,d}^{(s)}| &= x_{a,b,c,d}^{(s)}, & |\mathcal{O}_{a,b,c,d}^{(s)} - \mathbb{E}\mathcal{O}_{a,b,c,d}^{(s)}| &= y_{a,b,c,d}^{(s)} \text{ if } \mathbf{n}_s(a,b,c,d) \geq m\sqrt{n\log n} \\ |\mathcal{A}_{a,b,c,d}^{(s)} - \mathbb{E}\mathcal{A}_{a,b,c,d}^{(s)}| &= w_{a,b,c,d}^{(s)}, & |\mathcal{O}_{a,b,c,d}^{(s)} - \mathbb{E}\mathcal{O}_{a,b,c,d}^{(s)}| &= z_{a,b,c,d}^{(s)} \text{ if } \mathbf{n}_s(a,b,c,d) < m\sqrt{n\log n}. \end{aligned}$$

Conditioning on $\mathcal{E}_{\mathbf{x},\mathbf{y},\mathbf{w},\mathbf{z}}$, we note that each $\mathbf{n}_s(a,b,c,d) = O(mn)$ and so

$$\frac{1}{2}\|A - P_\tau A^T P_\tau^T\|_F^2 = \mathbb{E} \left(\frac{1}{2}\|A - P_\tau A^T P_\tau^T\|_F^2 \right) + O(m\sqrt{n\log n}).$$

A brief calculation yields

$$\begin{aligned} &\mathbb{E} \left(F_{\mathcal{A}}(A, B, \tau) + F_{\mathcal{O}}(A, B, \tau) \mid \mathcal{E}_{\mathbf{x},\mathbf{y},\mathbf{w},\mathbf{z}} \right) \\ &= (1 - \rho) \left[\sum_{i=1}^K \sum_{j=1}^K \binom{\epsilon_{i,j}^\tau}{2} \Lambda_{j,j} (1 - \Lambda_{j,j}) + \sum_{i=1}^K \sum_{j=1}^K \sum_{\ell > j}^K \epsilon_{i,j}^\tau \epsilon_{i,\ell}^\tau \Lambda_{j,\ell} (1 - \Lambda_{j,\ell}) + \sum_{i=1}^K \sum_{j=1}^K f_i^\tau \epsilon_{i,j}^\tau \Lambda_{i,j} (1 - \Lambda_{i,j}) \right. \\ &\quad + \sum_{i=1}^K \sum_{j > i}^K \sum_{\ell=1}^K \sum_{h=1}^K \epsilon_{i,h}^\tau \epsilon_{j,\ell}^\tau \Lambda_{h,\ell} (1 - \Lambda_{h,\ell}) + \sum_{i=1}^K \sum_{j > i}^K \sum_{\ell=1}^K f_i^\tau \epsilon_{j,\ell}^\tau \Lambda_{i,\ell} (1 - \Lambda_{i,\ell}) \\ &\quad \left. + \sum_{i=1}^K \sum_{j > i}^K \sum_{\ell=1}^K f_j^\tau \epsilon_{i,\ell}^\tau \Lambda_{\ell,j} (1 - \Lambda_{\ell,j}) \right] + O(m\sqrt{n\log n}). \end{aligned}$$

Note that

$$\begin{aligned}
& \sum_{i=1}^K \sum_{j=1}^K \binom{\epsilon_{i,j}^\tau}{2} \Lambda_{j,j} (1 - \Lambda_{j,j}) + \sum_{i=1}^K \sum_{j=1}^K \sum_{\ell > j}^K \epsilon_{i,j}^\tau \epsilon_{i,\ell}^\tau \Lambda_{j,\ell} (1 - \Lambda_{j,\ell}) + \sum_{i=1}^K \sum_{j=1}^K f_i^\tau \epsilon_{i,j}^\tau \Lambda_{i,j} (1 - \Lambda_{i,j}) \\
& + \sum_{i=1}^K \sum_{j > i}^K \sum_{\ell=1}^K \sum_{h=1}^K \epsilon_{i,h}^\tau \epsilon_{j,\ell}^\tau \Lambda_{h,\ell} (1 - \Lambda_{h,\ell}) + \sum_{i=1}^K \sum_{j > i}^K \sum_{\ell=1}^K f_i^\tau \epsilon_{j,\ell}^\tau \Lambda_{i,\ell} (1 - \Lambda_{i,\ell}) \\
& + \sum_{i=1}^K \sum_{j > i}^K \sum_{\ell=1}^K f_j^\tau \epsilon_{i,\ell}^\tau \Lambda_{\ell,j} (1 - \Lambda_{\ell,j}) \tag{26}
\end{aligned}$$

$$\begin{aligned}
& = \sum_{i=1}^K \sum_{j=1}^K \binom{\epsilon_{i,j}^\tau}{2} \Lambda_{i,i} (1 - \Lambda_{i,i}) + \sum_{i=1}^K \sum_{j=1}^K \sum_{\ell > j}^K \epsilon_{i,j}^\tau \epsilon_{i,\ell}^\tau \Lambda_{i,i} (1 - \Lambda_{i,i}) + \sum_{i=1}^K \sum_{j=1}^K f_i^\tau \epsilon_{i,j}^\tau \Lambda_{i,i} (1 - \Lambda_{i,i}) \\
& + \sum_{i=1}^K \sum_{j > i}^K \sum_{\ell=1}^K \sum_{h=1}^K \epsilon_{i,h}^\tau \epsilon_{j,\ell}^\tau \Lambda_{i,j} (1 - \Lambda_{i,j}) + \sum_{i=1}^K \sum_{j > i}^K \sum_{\ell=1}^K f_i^\tau \epsilon_{j,\ell}^\tau \Lambda_{i,j} (1 - \Lambda_{i,j}) \\
& + \sum_{i=1}^K \sum_{j > i}^K \sum_{\ell=1}^K f_j^\tau \epsilon_{i,\ell}^\tau \Lambda_{j,i} (1 - \Lambda_{j,i}), \tag{27}
\end{aligned}$$

and for any indices $i, j, k, \ell \in [K]$,

$$\Lambda_{i,j} (1 - \Lambda_{i,j}) + \Lambda_{k,\ell} (1 - \Lambda_{k,\ell}) \leq \Lambda_{i,j} (1 - \Lambda_{k,\ell}) + \Lambda_{k,\ell} (1 - \Lambda_{i,j}).$$

Writing $\mathbb{E} \left(F_{\mathcal{A}}(A, B, \tau) + F_{\mathcal{O}}(A, B, \tau) \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}} \right)$ as $(1 - \rho)(\frac{1}{2} \text{ of (26) } + \frac{1}{2} \text{ of (27)}) + O(m\sqrt{n \log n})$ yields

$$\mathbb{E} \left(F_{\mathcal{A}}(A, B, \tau) + F_{\mathcal{O}}(A, B, \tau) \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}} \right) \leq \frac{1 - \rho}{2} \mathbb{E} \left(\frac{1}{2} \|A - P_\tau A^T P_\tau^T\|_F^2 \right) + O(m\sqrt{n \log n}).$$

Conditioning on $\mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}}$, $F_{\mathcal{A}}(A, B, \tau) + F_{\mathcal{O}}(A, B, \tau)$ is the sum of

$$\frac{1}{2} \|A - P_\tau A^T P_\tau^T\|_F^2 = \mathbb{E} \left(\frac{1}{2} \|A - P_\tau A^T P_\tau^T\|_F^2 \right) + O(m\sqrt{n \log n})$$

independent Bernoulli random variables. Applying Hoeffding's inequality to $F := F_{\mathcal{A}}(A, B, \tau) + F_{\mathcal{O}}(A, B, \tau)$ yields that there exists constants $c_1 > 0$, c_2 , and c_3 , such that for n sufficiently large

$$\begin{aligned}
& \mathbb{P} \left(2F \geq \frac{1}{2} \|A - P_\tau A^T P_\tau^T\|_F^2 \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}} \right) \\
& = \mathbb{P} \left(2F - 2\mathbb{E}(F \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}}) \geq \frac{1}{2} \|A - P_\tau A^T P_\tau^T\|_F^2 - 2\mathbb{E}(F \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}}) \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}} \right) \\
& = \mathbb{P} \left(2F - 2\mathbb{E}(F \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}}) \geq \mathbb{E} \left(\frac{1}{2} \|A - P_\tau A^T P_\tau^T\|_F^2 \right) - 2\mathbb{E}(F \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}}) + O(m\sqrt{n \log n}) \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}} \right) \\
& \leq \mathbb{P} \left(F - \mathbb{E}(F \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}}) \geq \frac{\rho}{2} \mathbb{E} \left(\frac{1}{2} \|A - P_\tau A^T P_\tau^T\|_F^2 \right) + O(m\sqrt{n \log n}) \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}} \right) \\
& \leq \exp \left\{ -c_1 \rho^2 m n + c_2 \rho m \sqrt{n \log n} + c_3 m \log n \right\},
\end{aligned}$$

where the last inequality follows from $\mathbb{E}\left(\frac{1}{2}\|A - P_\tau A^T P_\tau^T\|_F^2\right) = \Omega(mn)$. Therefore, there exists a constant $\alpha > 0$ such that if $\rho > \sqrt{\alpha \frac{\log n}{n}}$ then

$$\mathbb{P}\left(2F \geq \frac{1}{2}\|A - P_\tau A^T P_\tau^T\|_F^2 \mid \mathcal{E}_{\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}}\right) \leq e^{-3m \log n}.$$

Combined with $\mathbb{P}(\mathcal{E}^c) \leq 24K^4 e^{-3m \log n}$, we have that, unconditionally,

$$\mathbb{P}\left(2F \geq \frac{1}{2}\|A - P_\tau A^T P_\tau^T\|_F^2\right) \leq (24K^4 + 1)e^{-3m \log n}.$$

Summing over τ and n yields that

$$\mathbb{P}(\exists \tau \in S_n \text{ with } X_{\tau, A, B} \leq -1) \leq (24K^4 + 1)e^{-3 \log n},$$

as desired. \square

A.4 Proof of Theorem 12

In this section, we will prove Theorem 12. We first restate the theorem.

Theorem 26. *With notation as above, let A and B be the adjacency matrices of ρ -SBM(K, \vec{n}, b, Λ) random graphs with K , and Λ fixed in n . Further assume there is an $\eta > 0$ such that $\Lambda \in [\eta, 1 - \eta]^{K \times K}$. Let $\{\tau_i\}_{i=1}^N$ be a collection of $N := \sum_i \lfloor \frac{n_i}{2} \rfloor$ disjoint within-block transpositions; i.e., if $\tau_i = k \leftrightarrow \ell$, then $b(k) = b(\ell) = b(\tau_i(k)) = b(\tau_i(\ell))$. For each $i = 1, 2, \dots, N$, let $E_{\tau_i, A, B}$ be the event $\{X_{\tau_i, A, B} \leq -1\}$. Let $X = \sum_i \mathbb{1}\{E_{\tau_i, A, B}\}$. There exists a constant $\beta > 0$ such that if $\rho \leq \sqrt{\beta \log n / n}$, then $\lim_{n \rightarrow \infty} \mathbb{P}(X = 0) = 0$.*

The proof of Theorem 12 will proceed as follows. We first show in Lemma 27 that for judiciously chosen β ,

$$\mathbb{E}(X) = \Omega\left(\frac{N}{n^{1/5} \sqrt{\log n}}\right).$$

In Lemmas 28–29, we show that

$$\text{Var}(X) = O\left(\frac{N^2}{\sqrt{n}}\right).$$

As X is a nonnegative integer-valued random variable, we apply the second moment method [3, Theorem 4.3.1] to derive

$$\mathbb{P}(X = 0) \leq \frac{\text{Var}(X)}{\mathbb{E}(X)^2} = O\left(\frac{\sqrt{\log n}}{n^{1/10}}\right) = o(1),$$

as desired. We now establish the supporting Lemmas 27–29. We begin by noting that if $\tau = i \leftrightarrow j$ is a within-block transposition, i.e., $b(i) = b(\tau(i)) = b(j) = b(\tau(j))$, then

$$\begin{aligned}
X_\tau &:= X_{\tau,A,B} = \|A - P_\tau B P_\tau^T\|_F^2 - \|A - B\|_F^2 \\
&= 2 \sum_{\substack{\ell,k \text{ s.t. } \tau(\ell) \neq \ell \\ \text{or } \tau(k) \neq k}} (A_{\ell,k} B_{\ell,k} - A_{\ell,k} B_{\tau(\ell),\tau(k)}) \\
&= 2 \sum_{k \neq i,j} (A_{i,k} B_{i,k} - A_{i,k} B_{j,k}) + 2 \sum_{k \neq i,j} (A_{j,k} B_{j,k} - A_{j,k} B_{i,k}) \\
&= 2 \sum_{k \neq i,j} (A_{i,k} - A_{j,k})(B_{i,k} - B_{j,k}). \tag{28}
\end{aligned}$$

For each $k \neq i, j$, the terms $X_\tau^{(k)} = X_{\tau,A,B}^{(k)} := 2(A_{i,k} - A_{j,k})(B_{i,k} - B_{j,k})$ in Eq. (28) are independent with mean

$$\mu_k := \mathbb{E}(2(A_{i,k} - A_{j,k})(B_{i,k} - B_{j,k})) = 4\Lambda_{b(i),b(k)}(1 - \Lambda_{b(i),b(k)})\rho,$$

and variance

$$\begin{aligned}
\sigma_k^2 &= \text{Var}(2(A_{i,k} - A_{j,k})(B_{i,k} - B_{j,k})) \\
&= 8(\Lambda_{b(i),b(k)}^2(1 - \Lambda_{b(i),b(k)})^2(1 - \rho)^2 + \Lambda_{b(i),b(k)}^2(1 - \Lambda_{b(i),b(k)})^2 + (1 - \Lambda_{b(i),b(k)})\Lambda_{b(i),b(k)}^3\rho \\
&\quad + \Lambda_{b(i),b(k)}(1 - \Lambda_{b(i),b(k)})^3\rho + \rho^2\Lambda_{b(i),b(k)}^2(1 - \Lambda_{b(i),b(k)})^2) - (4\Lambda_{b(i),b(k)}(1 - \Lambda_{b(i),b(k)})\rho)^2.
\end{aligned}$$

Note here that for all k ,

$$\lim_{\rho \rightarrow 0} \mu_k = 0, \quad \lim_{\rho \rightarrow 0} \sigma_k^2 = 16\Lambda_{b(i),b(k)}^2(1 - \Lambda_{b(i),b(k)})^2 > 0.$$

Next, note that

$$\begin{aligned}
\xi_k &:= \mathbb{E}\left(|2(A_{i,k} - A_{j,k})(B_{i,k} - B_{j,k}) - \mu_k|^3\right) \\
&= (2 + \mu_k)^3 2(\Lambda_{b(i),b(k)}^2(1 - \Lambda_{b(i),b(k)})^2(1 - \rho)^2) \\
&\quad + (2 - \mu_k)^3 2(\Lambda_{b(i),b(k)}(1 - \Lambda_{b(i),b(k)})(\Lambda_{b(i),b(k)} + \rho(1 - \Lambda_{b(i),b(k)}))(1 - \Lambda_{b(i),b(k)} + \rho\Lambda_{b(i),b(k)}) \\
&\quad + \mu_k^3(2\Lambda_{b(i),b(k)}^2 + 2(1 - \Lambda_{b(i),b(k)})^2)
\end{aligned}$$

we define

$$\mu_\tau = \sum_{k \neq i,j} \mu_k, \quad \sigma_\tau^2 = \sum_{k \neq i,j} \sigma_k^2, \quad \xi_\tau = \sum_{k \neq i,j} \xi_k.$$

The classic Berry-Esseen theorem [13, Theorem XVI.5.2] yields

$$\sup_x \left| \mathbb{P}\left(\frac{X_\tau - \mu_\tau}{\sigma_\tau} \leq x\right) - \Phi(x) \right| \leq \frac{6\xi_\tau}{\sigma_\tau^3},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Using the inequality [1, Eq. 7.1.13]

$$\frac{1}{x + \sqrt{x^2 + 2}} e^{-x^2} < \int_x^\infty e^{-t^2} dt \leq \frac{1}{x + \sqrt{x^2 + \frac{4}{\pi}}} e^{-x^2} \text{ for all } x \geq 0,$$

which is equivalent to

$$\frac{1}{\sqrt{\pi} \left(x + \sqrt{x^2 + 2} \right)} e^{-x^2} < 1 - \Phi \left(\sqrt{2}x \right) \leq \frac{1}{\sqrt{\pi} \left(x + \sqrt{x^2 + \frac{4}{\pi}} \right)} e^{-x^2} \text{ for all } x \geq 0, \quad (29)$$

we have that (with $x := \frac{1+\mu_\tau}{\sqrt{2}\sigma_\tau}$)

$$\mathbb{P}(X_\tau \leq -1) = \mathbb{P} \left(\frac{X_\tau - \mu_\tau}{\sigma_\tau} \leq -\sqrt{2}x \right) \leq \frac{1}{\sqrt{\pi} \left(x + \sqrt{x^2 + \frac{4}{\pi}} \right)} e^{-x^2} + \frac{6\xi_\tau}{\sigma_\tau^3}, \quad (30)$$

and

$$\mathbb{P}(X_\tau \leq -1) = \mathbb{P} \left(\frac{X_\tau - \mu_\tau}{\sigma_\tau} \leq -\sqrt{2}x \right) \geq \frac{1}{\sqrt{\pi} \left(x + \sqrt{x^2 + 2} \right)} e^{-x^2} - \frac{6\xi_\tau}{\sigma_\tau^3}. \quad (31)$$

Lemma 27. *With notation as above, there exists a constant β such that if $\rho \leq \sqrt{\frac{\beta \log n}{n}}$, then $\mathbb{P}(X_\tau \leq -1) = \Omega \left(\frac{1}{n^{1/4} \sqrt{\log n}} \right)$*

Proof. We first note that there exists a constant $c_1 > 0$ such that for n sufficiently large, $\sigma_k^2 \geq c_1$ holds for all k . Therefore, if $\rho \leq \sqrt{\frac{\beta \log n}{n}}$ then

$$x = \frac{1 + \mu_\tau}{\sqrt{2}\sigma_\tau} \leq \frac{1 + \rho n}{\sqrt{2}c_1 n} \leq \frac{1 + \sqrt{\beta n \log n}}{\sqrt{2}c_1 n} = \sqrt{\frac{\beta \log n}{2c_1}} + \frac{1}{\sqrt{2}c_1 n}, \quad (32)$$

and β can be chosen so that $\rho \leq \sqrt{\frac{\beta \log n}{n}}$ implies that $x \leq \sqrt{\frac{\log n}{5}}$. With this choice of β , The lower bound in Eq. (31) is then bounded by

$$\frac{1}{\sqrt{\pi} \left(x + \sqrt{x^2 + 2} \right)} e^{-x^2} - \frac{6\xi_\tau}{\sigma_\tau^3} \geq \frac{\exp \left\{ -\frac{\log n}{5} \right\}}{\sqrt{\pi} \left(\sqrt{\frac{\log n}{5}} + \sqrt{\frac{\log n}{5}} + 2 \right)} - \Theta \left(n^{-1/2} \right) = \Omega \left(\frac{1}{n^{1/5} \sqrt{\log n}} \right),$$

as desired. □

Recalling that $E_{\tau_i} = E_{\tau_i, A, B}$ is the event $\{X_{\tau_i} \leq -1\}$, Lemma 27 is equivalent to

$$\mathbb{E}(\mathbb{1}\{E_{\tau_i}\}) = \mathbb{P}(E_{\tau_i}) = \mathbb{P}(X_{\tau_i} \leq -1) = \Omega \left(\frac{1}{n^{1/5} \sqrt{\log n}} \right).$$

It follows immediately that (where N is as defined in Theorem 12)

$$\mathbb{E}(X) = \Omega \left(\frac{N}{n^{1/5} \sqrt{\log n}} \right).$$

We now turn our attention to bounding $\text{Var}(X)$.

Lemma 28. *With notation as above and assumptions as in Theorem 12, let $\tau_1 = i \leftrightarrow j$ and $\tau_2 = h \leftrightarrow \ell$ be two disjoint, within-block transpositions. There exists constants $C_1 > 0$ and $C_2 > 0$,*

$$\text{Cov}(\mathbb{1}\{E_{\tau_i}\}, \mathbb{1}\{E_{\tau_j}\}) \leq C_2 \left(\exp\{-C_1 \rho^2 n\} + \Theta\left(\frac{1}{\sqrt{n}}\right) \right) \Theta\left(\frac{1}{\sqrt{n}}\right)$$

Proof. Note that

$$\text{Cov}(\mathbb{1}\{E_{\tau_1}\}, \mathbb{1}\{E_{\tau_2}\}) = \mathbb{P}(X_{\tau_1} \leq -1, X_{\tau_2} \leq -1) - \mathbb{P}(X_{\tau_1} \leq -1)\mathbb{P}(X_{\tau_2} \leq -1).$$

Next, observe that X_{τ_1} and X_{τ_2} are each the sum of $n - 2$ independent terms ($\{X_{\tau_1}^{(k)}\}_{k \neq i, j}$ and $\{X_{\tau_2}^{(k)}\}_{k \neq h, \ell}$ resp.) which are collectively independent except for the four terms

$$\begin{aligned} X_{\tau_1}^{(h)} &= (A_{i, h} - A_{j, h})(B_{i, h} - B_{j, h}), \quad X_{\tau_1}^{(\ell)} = (A_{i, \ell} - A_{j, \ell})(B_{i, \ell} - B_{j, \ell}), \\ X_{\tau_2}^{(i)} &= (A_{h, i} - A_{\ell, i})(B_{h, i} - B_{\ell, i}), \quad X_{\tau_2}^{(j)} = (A_{h, j} - A_{\ell, j})(B_{h, j} - B_{\ell, j}). \end{aligned}$$

Let $\tilde{X}_{\tau_2} = X_{\tau_2} - X_{\tau_2}^{(i)} - X_{\tau_2}^{(j)}$, so that X_{τ_1} and \tilde{X}_{τ_2} are independent. Noting that $|\tilde{X}_{\tau_2} - X_{\tau_2}| \leq 2$, we have

$$\mathbb{P}(X_{\tau_1} \leq -1, X_{\tau_2} \leq -1) \leq \mathbb{P}(X_{\tau_1} \leq -1, \tilde{X}_{\tau_2} \leq 1) = \mathbb{P}(X_{\tau_1} \leq -1)\mathbb{P}(\tilde{X}_{\tau_2} \leq 1).$$

Therefore,

$$\begin{aligned} \text{Cov}(\mathbb{1}\{E_{\tau_1}\}, \mathbb{1}\{E_{\tau_2}\}) &\leq \mathbb{P}(X_{\tau_1} \leq -1) \left(\mathbb{P}(\tilde{X}_{\tau_2} \leq 1) - \mathbb{P}(X_{\tau_2} \leq -1) \right) \\ &\leq \mathbb{P}(X_{\tau_1} \leq -1) \left(\mathbb{P}(\tilde{X}_{\tau_2} \leq 1) - \mathbb{P}(\tilde{X}_{\tau_2} \leq -3) \right) \end{aligned} \quad (33)$$

As in Eq. (32), as there is a constant $c_2 > 0$ such that $\sigma_k^2 \leq c_2$ for all k , we have that

$$x_1 := \frac{1 + \mu_{\tau_1}}{\sqrt{2}\sigma_{\tau_1}} \geq \frac{1 + 4\eta(1 - \eta)\rho n}{\sqrt{2}c_2 n} \geq \frac{4\eta(1 - \eta)\rho n}{\sqrt{2}c_2 n} = \frac{4\eta(1 - \eta)\rho\sqrt{n}}{\sqrt{2}c_2}, \quad (34)$$

and from Eq. (30) we have that

$$\begin{aligned} \mathbb{P}(X_{\tau_1} \leq -1) &\leq \frac{\exp\left\{-\left(\frac{4\eta(1-\eta)\rho\sqrt{n}}{\sqrt{2}c_2}\right)^2\right\}}{2} + \Theta\left(\frac{1}{\sqrt{n}}\right) \\ &= \frac{1}{2}\exp\{-C_1 \rho^2 n\} + \Theta\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \quad (35)$$

for a constant $C_1 > 0$. Define $\tilde{\mu}_{\tau_2} = \mathbb{E}(\tilde{X}_{\tau_2})$, and $\tilde{\sigma}_{\tau_2}^2 = \text{Var}(\tilde{X}_{\tau_2})$. We have that

$$\begin{aligned} \mathbb{P}\left(\tilde{X}_{\tau_2} \in [-2, 1]\right) &= \mathbb{P}\left(\frac{\tilde{X}_{\tau_2} - \tilde{\mu}_{\tau_2}}{\tilde{\sigma}_{\tau_2}} \leq \frac{1 - \tilde{\mu}_{\tau_2}}{\tilde{\sigma}_{\tau_2}}\right) - \mathbb{P}\left(\frac{\tilde{X}_{\tau_2} - \tilde{\mu}_{\tau_2}}{\tilde{\sigma}_{\tau_2}} \leq \frac{-3 - \tilde{\mu}_{\tau_2}}{\tilde{\sigma}_{\tau_2}}\right) \\ &= \Phi\left(\frac{1 - \tilde{\mu}_{\tau_2}}{\tilde{\sigma}_{\tau_2}}\right) - \Phi\left(\frac{-3 - \tilde{\mu}_{\tau_2}}{\tilde{\sigma}_{\tau_2}}\right) + \Theta\left(\frac{1}{\sqrt{n-2}}\right) \\ &\leq \frac{4}{\sqrt{2\pi}\tilde{\sigma}_{\tau_2}} + \Theta\left(\frac{1}{\sqrt{n-2}}\right) = \Theta\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \quad (36)$$

Combining Eq. (35) and (36) yields the desired result. \square

We now combine Lemmas 27 and 28 to bound $\text{Var}(X_\tau)$ where τ is a within-block transposition.

Lemma 29. *With notation as above and assumptions as in Theorem 12, let τ be a within-block transposition. We have that*

$$\text{Var}(X_\tau) = O\left(\frac{N^2}{\sqrt{n}}\right)$$

Proof. Recall from Eq. (35) that there exists a constant $C_1 > 0$ such that

$$\mathbb{P}(X_\tau \leq -1) \leq \frac{1}{2} \exp\{-C_1 \rho^2 n\} + \Theta\left(\frac{1}{\sqrt{n}}\right),$$

so that

$$\text{Var}(\mathbb{1}\{X_\tau \leq -1\}) = (1 - \mathbb{P}(X_\tau \leq -1))\mathbb{P}(X_\tau \leq -1) \leq \frac{1}{2} \exp\{-C_1 \rho^2 n\} + \Theta\left(\frac{1}{\sqrt{n}}\right).$$

Combined with Lemma 28,

$$\text{Var}(X) \leq \frac{N}{2} e^{-C_1 \rho^2 n} + \Theta\left(\frac{N}{\sqrt{n}}\right) + N^2 C_2 \left(e^{-C_1 \rho^2 n} + \Theta\left(\frac{1}{\sqrt{n}}\right)\right) \Theta\left(\frac{1}{\sqrt{n}}\right) = O\left(\frac{N^2}{\sqrt{n}}\right)$$

as desired. \square

A.5 Proof of Theorem 15

The key to the proof of Theorem 15 is the following consequence of Theorem 10: If $(G_1, G_2) \sim \rho$ -correlated SBM(K, \vec{n}, b, Λ) with $\rho = \omega(\sqrt{\log n/n})$ and respective adjacency matrices A and B , then $\text{argmin}_{P \in \Pi(n)} \|AP - PB\|_F = \{I_n\}$ with high probability.

Proposition 30. *Let $(G_1, \sigma(G_2)) \sim \sigma, \rho$ -correlated SBM(K, \vec{n}, b, Λ) with $\rho = \omega(\sqrt{\log n/n})$. We have that*

$$\mathbb{P}[G_{\sigma(G_2) \rightarrow G_1} \neq G_2] = O(e^{-3 \log n}).$$

Proof. To ease notation, define

$$\begin{aligned} \mathbb{P}(x, y, z) &= \mathbb{P}[(G_1, G_2, G_{\sigma(G_2) \rightarrow G_1}) = (x, y, z)], \\ \mathbb{P}(x, z) &= \mathbb{P}[(G_1, G_{\sigma(G_2) \rightarrow G_1}) = (x, z)], \\ \mathbb{P}(x, y) &= \mathbb{P}[(G_1, G_2) = (x, y)]. \end{aligned}$$

Note that

$$\begin{aligned} \mathbb{P}[G_{\sigma(G_2) \rightarrow G_1} \neq G_2] &= \sum_{\substack{(x,y,z) \\ \text{s.t. } z \neq y}} \mathbb{P}(x, y, z) = \sum_{\substack{(x,y,z) \\ \text{s.t. } z \neq y}} \mathbb{P}(x, y) \frac{\mathbb{1}\{z \in P_{x,y}^*(y)\}}{|P_{x,y}^*(y)|} \\ &= \sum_{\substack{(x,y,z) \text{ s.t. } z \neq y, \\ z \in P_{x,y}^*(y)}} \frac{\mathbb{P}(x, y)}{|P_{x,y}^*(y)|} = \sum_{x,y} \sum_{\substack{z \in P_{x,y}^*(y) \\ \text{s.t. } z \neq y}} \frac{\mathbb{P}(x, y)}{|P_{x,y}^*(y)|} \leq \sum_{\substack{x,y \text{ s.t.} \\ P_{x,y}^*(y) \neq \{y\}}} \mathbb{P}(x, y). \end{aligned}$$

Theorem 10 implies that

$$\sum_{\substack{x,y \text{ s.t.} \\ P_{x,y}^* \neq \{I\}}} \mathbb{P}(x, y) = O(e^{-3 \log n}).$$

As $P_{x,y}^*(y) \neq \{y\} \Rightarrow P_{x,y}^* \neq \{I\}$, this implies

$$\sum_{\substack{x,y \text{ s.t.} \\ P_{x,y}^*(y) \neq \{y\}}} \mathbb{P}(x, y) \leq \sum_{\substack{x,y \text{ s.t.} \\ P_{x,y}^* \neq \{I\}}} \mathbb{P}(x, y) = O(e^{-3 \log n}),$$

as desired. \square

Theorem 15 is then a straightforward application of Fano's inequality, which yields that

$$H[G_{\sigma(G_2) \rightarrow G_1} | G_2] = o(1), \text{ and } H[G_2 | G_{\sigma(G_2) \rightarrow G_1}] = o(1).$$

By the chain rule for entropy, we have that

$$\begin{aligned} H[G_1, G_2, G_{\sigma(G_2) \rightarrow G_1}] &= H[G_1] + H[G_{\sigma(G_2) \rightarrow G_1} | G_1] + \underbrace{H[G_2 | G_{\sigma(G_2) \rightarrow G_1}, G_1]}_{=o(1)} \\ &= H[G_1] + H[G_2 | G_1] + \underbrace{H[G_{\sigma(G_2) \rightarrow G_1} | G_2, G_1]}_{=o(1)}. \end{aligned}$$

From this we have that $H[G_2 | G_1] = H[G_{\sigma(G_2) \rightarrow G_1} | G_1] + o(1)$. Finally,

$$\begin{aligned} H[G_1, G_2, G_{\sigma(G_2) \rightarrow G_1}] &= H[G_2] + \underbrace{H[G_{\sigma(G_2) \rightarrow G_1} | G_2]}_{=o(1)} + H[G_1 | G_{\sigma(G_2) \rightarrow G_1}, G_2] \\ &= H[G_{\sigma(G_2) \rightarrow G_1}] + \underbrace{H[G_2 | G_{\sigma(G_2) \rightarrow G_1}]}_{=o(1)} + H[G_1 | G_{\sigma(G_2) \rightarrow G_1}, G_2], \end{aligned}$$

so that $H[G_{\sigma(G_2) \rightarrow G_1}] = H[G_2] + o(1)$. Combined, this yields that

$$I(G_1; G_2) - I(G_1; G_{\sigma(G_2) \rightarrow G_1}) = o(1),$$

as desired.

References

- [1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Number 55. Courier Corporation, 1964.
- [2] E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Advances in Neural Information Processing Systems*, 26:692–700, 2013.

- [3] N. Alon and J. H. Spencer. *The probabilistic method*. John Wiley & Sons, 2015.
- [4] D. Asta and C. R. Shalizi. Geometric network comparison. *arXiv preprint arXiv:1411.1350*, 2014.
- [5] L. Babai. Graph isomorphism in quasipolynomial time. *arXiv preprint arXiv:1512.03547*, 2015.
- [6] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.
- [7] P. J. Carrington, J. Scott, and S. Wasserman. *Models and methods in social network analysis*, volume 28. Cambridge university press, 2005.
- [8] S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2014.
- [9] L. Chen, J.T. Vogelstein, V. Lyzinski, and C. Priebe. A joint graph inference case study: the c.elegans chemical and electrical connectomes. *arXiv preprint, submitted*, 2015.
- [10] D. Choi and P. J. Wolfe. Co-clustering separately exchangeable network data. *The Annals of Statistics*, 42(1):29–63, 2014.
- [11] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298, 2004.
- [12] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [13] W. Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 2008.
- [14] M. Fiori, P. Sprechmann, J. Vogelstein, P. Mus, and G. Sapiro. Robust multimodal graph matching: Sparse coding meets graph matching. *Advances in Neural Information Processing Systems*, pages 127–135, 2013.
- [15] D. E. Fishkind, D. L. Sussman, M. Tang, J. T. Vogelstein, and C. E. Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.
- [16] D.E. Fishkind, S. Adali, and C.E. Priebe. Seeded graph matching. *arXiv preprint arXiv:1209.0367*, 2012.
- [17] P. Foggia, G. Percannella, and M. Vento. Graph matching and learning in pattern recognition in the last 10 years. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(01):1450001, 2014.
- [18] C. Fraley and A. E. Raftery. Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16(2):297–306, 1999.

- [19] W. R. Gray, J. A. Bogovic, J. T. Vogelstein, B. A. Landman, J. L. Prince, and R. J. Vogelstein. Magnetic resonance connectome automated pipeline: an overview. *Pulse, IEEE*, 3(2):42–48, 2012.
- [20] D. Hardoon, S. Szedmak, and J. S. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.
- [21] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.
- [22] A. Kandel, H. Bunke, and M. Last. *Applied Graph Theory in Computer Vision and Pattern Recognition*, volume 1. Springer, 2007.
- [23] J. H. Kim, B. Sudakov, and V. H. Vu. On the asymmetry of random regular graphs and random graphs. *Random Structures and Algorithms*, 21:216–224, 2002.
- [24] N. H. Lee and C. E. Priebe. A latent process model for time series of attributed random graphs. *Statistical inference for stochastic processes*, 14(3):231–253, 2011.
- [25] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.
- [26] V. Lyzinski, S. Adali, J. T. Vogelstein, Y. Park, and C. E. Priebe. Seeded graph matching via joint optimization of fidelity and commensurability. *arXiv preprint arXiv:1401.3813*, 2014.
- [27] V. Lyzinski, D. Fishkind, M. Fiori, J.T. Vogelstein, C.E. Priebe, and G. Sapiro. Graph matching: Relax at your own risk. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, In press, 2015.
- [28] V. Lyzinski, D.E. Fishkind, and C.E. Priebe. Seeded graph matching for correlated Erdos-Renyi graphs. *Journal of Machine Learning Research*, 15:3513–3540, 2014.
- [29] V. Lyzinski, Y. Park, C. E. Priebe, and M. Trosset. Fast embedding for jofc using the raw stress criterion. *arXiv preprint, arXiv:1502.03391*, 2015.
- [30] V. Lyzinski, D. L. Sussman, D. E. Fishkind, H. Pao, L. Chen, J. T. Vogelstein, Y. Park, and C. E. Priebe. Spectral clustering for divide-and-conquer graph matching. *Parallel Computing*, 47:70–87, 2015.
- [31] V. Lyzinski, D. L. Sussman, M. Tang, A. Athreya, and C. E. Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905–2922, 2014.
- [32] C. L. M. Nickel. *Random dot product graphs: A model for social networks*. PhD thesis, Johns Hopkins University, 2006.
- [33] E. Onaran, S. Garg, and E. Erkip. Optimal de-anonymization in random graphs with community structure. *arXiv preprint arXiv:1602.01409*, 2016.

- [34] C. E. Priebe, N. H. Lee, Y. Park, and M. Tang. Attribute fusion in a latent process model for time series of graphs. In *The 2011 IEEE Workshop on Statistical Signal Processing (SSP2011)*, 2011.
- [35] C.E. Priebe, D.J. Marchette, Z. Ma, and S. Adali. Manifold matching: joint optimization of fidelity and commensurability. *Brazilian Journal of Probability and Statistics*, 27(3):377400, 2013.
- [36] T. Qin and K. Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. *Advances in Neural Information Processing Systems*, 2013.
- [37] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
- [38] J. Richiardi, H. Eryilmaz, S. Schwartz, P. Vuilleumier, and D. Van De Ville. Decoding brain states from fmri connectivity graphs. *Neuroimage*, 56(2):616–626, 2011.
- [39] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, 39:1878–1915, 2011.
- [40] C. Shen and C. E. Priebe. Manifold matching using shortest-path distance and joint neighborhood selection. *arXiv preprint arXiv:1412.4098*, 2014.
- [41] M. Sun and C. E. Priebe. Efficiency investigation of manifold matching for text document classification. *Pattern Recognition Letters*, 34(11):1263–1269, 2013.
- [42] M. Sun, C. E. Priebe, and M. Tang. Generalized canonical correlation analysis for disparate data fusion. *Pattern Recognition Letters*, 34(2):194–200, 2013.
- [43] D. L. Sussman, M. Tang, D. E. Fishkind, and C. E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.
- [44] M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, and C. E. Priebe. A nonparametric two-sample hypothesis testing problem for random dot product graphs. *Bernoulli*, To appear, 2015.
- [45] M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, and C. E. Priebe. A semiparametric two-sample hypothesis testing problem for random dot product graphs. *Journal of Computational and Graphical Statistics*, Accepted for publication, 2016.
- [46] M. Tang, Y. Park, N. H. Lee, and C. E. Priebe. Attribute fusion in a latent process model for time series of graphs. *IEEE Transactions on Signal Processing*, 61(7):1721–1732, 2013.
- [47] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii. Structural properties of the caenorhabditis elegans neuronal network. *PLoS computational biology*, 7(2):e1001066, 2011.
- [48] J. T. Vogelstein and C. E. Priebe. Shuffled graph classification: Theory and connectome applications. *Journal of Classification*, 32(1):3–20, 2015.

- [49] J.T. Vogelstein, J.M. Conroy, V. Lyzinski, L.J. Podrazik, S.G. Kratzer, E.T. Harley, D.E. Fishkind, R.J. Vogelstein, and C.E. Priebe. Fast Approximate Quadratic Programming for Graph Matching. *PLoS ONE*, 10(04), 2014.
- [50] H. Wang, M. Tang, Y. Park, and C. E. Priebe. Locality statistics for anomaly detection in time series of graphs. *Signal Processing, IEEE Transactions on*, 62(3):703–717, 2014.
- [51] Y. J. Wang and G. Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.
- [52] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode *caenorhabditis elegans*. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 314(1165):1–340, 1986.
- [53] P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. arXiv preprint at <http://arxiv.org/abs/1309/5936>, 2013.
- [54] S. Young and E. Scheinerman. Random dot product graph models for social networks. In *Proceedings of the 5th international conference on algorithms and models for the web-graph*, pages 138–149, 2007.
- [55] M. Zaslavskiy, F. Bach, and J.P. Vert. A path following algorithm for the graph matching problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2227–2242, 2009.